Bayesian Modelling in Genetic Association Studies

Jemma Walker

June 2012

Thesis submitted to the University of London in fulfilment of the requirements for the Doctorate of Philosophy

Faculty of Epidemiology and Population Health

London School of Hygiene and Tropical Medicine

Keppel Street

London WC1E 7HT

I, Jemma Louise Walker, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Firstly, I would like to thank my supervisors Dr. Juliet Chapman and Professor John Whittaker for all their invaluable help, ideas, suggestions, expertise and time. I have learnt a lot throughout this experience.

I would also like to extend my gratitude to Dr. Claudio Verzilli, Professor Aroon Hingorani, Dr. Fotios Drenos, Professor Tim Vyse and Dr. David Morris for their input and advice. Thanks also to Professor Bianca DeStavola, Dr. Dorothea Nitsch and ,Dr. Branwen Hennig for their helpful comments.

I am grateful to all my friends for their support and patience throughout this work ... I am looking forward to getting my social life back and seeing you all. Special thanks to my LSHTM friends for all the coffee and lunch breaks. Particularly, for all the encouragement when it was most needed to Juliet, Emily, Zoe, Kate, Cat, Caroline and Caroline.

I greatly appreciate funding from the Wellcome Trust, under which this work was completed.

Finally, I would like to thank my family for supporting me in whatever I do.

CONTENTS

CHAPTER

1

ABSTRACT

Bayesian Model Selection Approaches are flexible methods that can be utilised to investigate Genetic Association studies in greater detail; enabling us to more accurately pin-point locations of disease genes in complex regions such as the MHC, as well as investigate possible causal pathways between genes, disease and intermediate phenotypes. This thesis is split into two distinct parts. The first uses a Bayesian Multivariate Adaptive Regression Spline Model to search across many highly correlated variants to try to determine which are likely to be the truly causal variants within complex genetic regions and also how each of these variants influences disease status. Specifically, I consider the role of genetic variants within the MHC region on SLE. The second part of the thesis aims to model possible disease pathways between genes, disease, intermediate phenotypes and environmental factors using Bayesian Networks, in particular focussing upon Coronary Heart disease and numerous blood biomarkers and related genes.

Bayesian Multivariate Adaptive Regression Spline Model

Genetic association studies have the problem that often many genotypes in strong linkage disequilibrium (LD) are found to be associated with the outcome of interest. This makes it difficult to establish the actual SNP responsible.

The aim of this part of the thesis is to investigate Bayesian variable selection methods in regions of high LD. In particular, to investigate SNPs in the major histocompatibility complex (MHC) region associated with systematic lupus erythematosus (SLE). Past studies have found several SNPs in this region to be highly associated with SLE but these SNPs are in high LD with one another.

It is desirable to search over all possible regression models in order to find those SNPs that are most important in the prediction of SLE. The Bayesian Multivariate Adapative Regression Splines (BMARS) model used should automatically correct for nearby associated SNPs, and only those directly associated should be included in the model. The BMARS approach will also automatically select the most appropriate disease model for each directly associated variant.

It was found that there appear to be 3 separate SNP signals in the MHC region that show association with SLE. The rest of the associations found using simple Frequentist tests are likely to be due to LD with the true signal.

Bayesian Networks for Genetic Association Studies

Coronary Heart Disease (CHD) is one of many diseases that result from complicated relationships between both genetic and environmental factors. Identifying causal factors and developing new treatments that target these factors is very difficult. Changes in intermediate phenotypes, or biomarkers, could suggest potential causal pathways, although these have a tendency to group amongst those patients with higher risk of CHD making to difficult to distinguish independent causal relationships. I aim to model disease pathways allowing for intermediate phenotypes as well as genetic and environmental factors.

Statistical methodology was developed using directed acyclic graphs (DAGs). Disease outcomes, genes, intermediate phenotypes and possible explanatory variables were represented as nodes in a DAG. Possible models were investigated using Bayesian regression models, based upon the underlying DAG, in a reversible jump MCMC framework. Modelling the data this way allows us to distinguish between direct and indirect effects as well as explore possible directionality of relationships. Since different DAGs can belong to the same equivalence class, some directions of association may become indistinguishable and I am interested in the implications of this.

I investigated the integrated associations of genotypes with multiple blood biomarkers linked to CHD risk, focusing particularly on relationships between APOE, CETP and APOB genotypes; HDL- and LDL- cholesterol, triglycerides, C-reactive protein, fibrogen and apolipoproteins A and B.

Overview

I will begin by introducing the topics of genetics, statistics and directed acyclic graphs with a background on each (Chapters 2, 3 and 4 respectively). Chapter 5 will then detail the analysis and results of the BMARS model. The analysis and results of Bayesian networks for genetic association studies will then be covered in Chapter 6.

CHAPTER

2

BACKGROUND TO GENETICS

The material in sections 2.1 - 2.4 is all referenced from Molecular Biology of The Cell by Alberts, Johnson, Lewis, Raff, Roberts & Walter [1], An Introduction to Genetic Analysis by Griffiths, Miller, Suzuki, Lewontin & Gelbart [2] and Essentials Of Medical Genomics by Brown [3].

Genetics is the key to heredity and variation in living organisms. Genetic information is stored in the nucleus of most cells of an organism. This information is both copied and passed onto offspring (through replication of DNA); and translated into proteins (used for different functions within the organism). These processes are described in more detail below.

2.1 DNA, Chromosomes and Genes

Genetic information is transmitted and stored as deoxyribonucleic acid (DNA). DNA is made up of strings of polymers called nucleotides, or bases. Nucleotides come in

four types: adenine (A), cytosine (C), guanine (G), and thymine (T). These strings of nucleotides form a double helix and the nucleotides on each strand pair with the one opposite via hydrogen bonds. A pairs with T, and C pairs with G. Therefore, one strand completely defines the other and just one of the complementary bases will define the nucleotide type at any given position along the DNA chain.

Figure 2.1: DNA double helix [4]

The **genome** is the complete set of genetic information contained in the DNA of an organism. A **gene** is a unit of heredity which carries information from one generation to the next.

This information is stored on very long units in which DNA is packaged called **chromosomes.** Chromosomes are all packed together very tightly in the cell nucleus. Other than gamete cells (see below) humans have 46 chromosomes in each cell nucleus- one pair of sex chromosomes and 22 pairs of autosomes (non sex chromosomes) . In reproduction, each parent provides an offspring with one chromosome of each of their 23 pairs. Gamete cells are sperm and egg cells and have 23 single chromosomes (haploid), rather than 23 pairs (diploid). In reproduction, gamete cells fuse with gamete cells from the opposite sex.

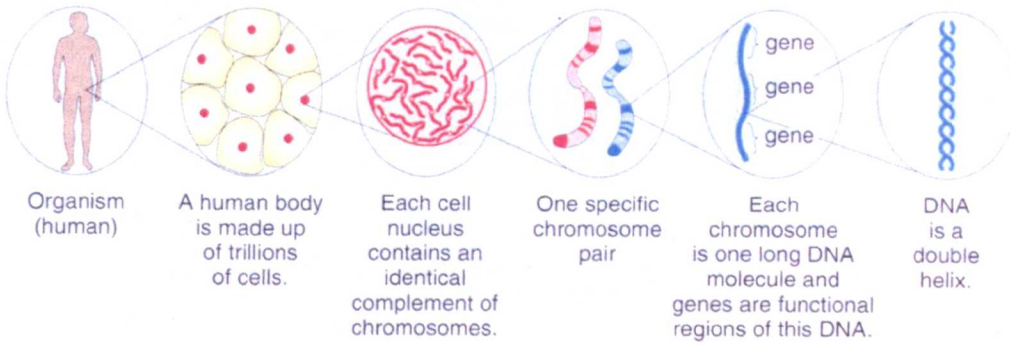


Figure 2.2: Successive enlargements of an organism to focus on the genetic material [2]

2.2 Replication, Transcription & Translation

When DNA replicates the double helix structure unwinds and splits into two strands. Each single strand of DNA acts as a template for the production of a new strand with complementary bases. These new strands pair with their templates and form two double stranded helix molecules of DNA identical to the original (barring any novel genetic variation).

As mentioned before, genes are functional regions of DNA. The genetic information stored in DNA can also be translated into **protein** . Producing proteins from information in a DNA gene is a two step process. The first step is called transcription, and involves the synthesis of a riboneuclic acid (RNA) chain that is complementary to one of the strands of DNA. RNA is similar to DNA, and is made up of a string of bases: A, C, G & Uracil (U), instead of T. RNA has the same complementary bases as DNA (A pairs with U). To transcribe the information in DNA, the double helix separates, and one of the strands acts as the template to form a complementary strand of RNA.

The second step in producing a protein from DNA is called translation, and involves using the information in RNA to produce protein. Proteins are responsible for many functions in the cell. For example, they act as enzymes or structural components, and they are essential in building muscle, skin, bone and blood.

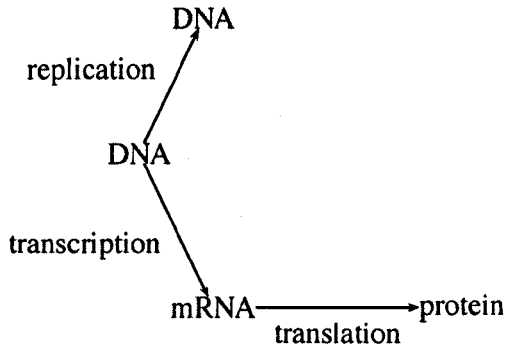


Figure 2.3: Steps in Producing DNA and Proteins

2.3 Mendelian Laws of Inheritance

A **locus** is a particular position along the genome and can refer to a single base or longer region. Genes relating to a particular trait (eg. eye colour) are located at the same locus on the same chromosome in each individual. Every individual has two of every chromosome (one from each parent) and therefore two of each locus. **Alleles** are different possible forms of the locus. Each individual can only have two alleles for each locus. These can vary between individuals. A **phenotype** is a detectable outward manifestation of a gene. For example, the gene for the phenotype eye colour has alleles that result in blue, brown or green eyes.

As mentioned above, offspring inherit one chromosome of a pair from each parent. Mendel's principle of segregation states that the allele inherited from each parent for each characteristic is random with equal probability. In the diagram 2.4 below, two individuals have alleles M1, M2 and M3, M4 respectively for marker 1 and 2 for parent 1 and 2. These alleles could each be A,T,C or G. These alleles separate from each other, and then combine with another allele from the other parent. The combination of these alleles is random with equal probability. The possible allele combinations from these two parents are M1/M3, M1/M4, M2/M3 or M2/M4. Each allele combination occurs with 25% probability, assuming the alleles are not linked.

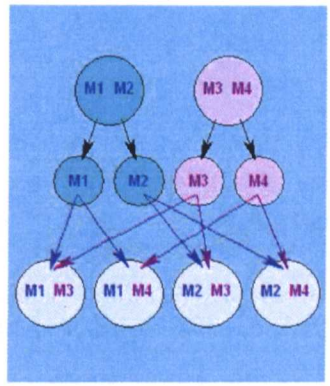


Figure 2.4: Different Allele Combinations from Two Parents

The combination of unordered alleles at a particular locus is known as a **genotype.** If an individual has two of the same allele for a particular trait then it is known as **homozygous.** However, if an individual has two different alleles then it is **heterozygous.** For example, in the stretch of DNA shown below in Table 2.1, shows the two chromosomes one above the other, for a segment of DNA made up of ten base pairs. The third base pair is highlighted and two different alleles are possible at this locus (namely A and T). The individual is heterozygous at this locus.

T	G	**A**	A	A	G	A	C	C	A
C	C	**T**	G	T	C	A	G	C	T

Table 2.1: Example of Genotypes

Suppose that only two alleles (A and T) are possible at this locus. The genotype of the highlighted locus is A/T, but could be A/A (homozygous) or T/T (homozygous) in a different individual. Note that a genotype is unordered so that A/T and T/A are equivalent.

2.3.1 Hardy-Weinberg Equilibrium

In generalised terms, if we have a single locus with two possible alleles, A_1 with probability p, and A_2 with probability (1-p) in a population, the expected frequencies are

shown in Table 2.2 below:

	Mother	
Father	**A_1(p)**	**A_2(1-p)**
A_1(p)	$A_1A_1(p^2)$	$A_1A_2(p(1-p))$
A_2(1-p)	$A_1A_2(p(1-p))$	$A_2A_2((1-p)^2)$

Table 2.2: Frequencies of Alleles Inherited under Hardy-Weinberg Equilibrium

These frequencies are expected under the assumption that individuals in a population are randomly mating and therefore diploid genotype (genotypes with information about both alleles rather than genotypes coded as 0,1,2 as described later) frequencies should only depend on allele frequencies in the population. For example, if the allele frequency of A_1 is 0.9 and that of A_2 is 0.1 then the expected genotype frequency of A_1/A_1 is 0.81, A_1/A_2 is 0.18, and A_2/A_2 is 0.01. If the observed genotype frequencies (calculated as described above) follow these probability expectations then this is known as **Hardy-Weinberg equilibrium** (HWE). HWE is usually tested by using a chi-squared test, comparing the observed genotype frequencies with those expected under HWE. If the genotype frequency deviates from HWE for a particular locus then this may be due to several reasons including genotyping errors, sampling variation, and non-random mating. Therefore, HWE tests are often used as a form of quality control test for genotype data and to check whether assumptions made in subsequent analyses are reasonable.

2.4 Haplotypes and Recombination

A **haplotype** is a combination of alleles transmitted together at multiple loci on the same chromosome. If there are two different alleles possible at each of two loci (for example alleles A_1 and A_2 at locus 1 and B_1 and B_2 at locus 2) then there are four possible haplotypes (A_1-B_1, A_1-B_2, A_2-B_1 and A_2-B_2). Each individual will have two haplotypes (one on each strand).

Occasionally when chromosomes pair during meiosis (the process of forming gametes), the chromosomes exchange segments/strands during a process called crossing-over. This results in new combinations of alleles called **recombinants** . Recombinants can cause offspring to have haplotypes not seen in the parents. Figure 2.5 demonstrates crossing over of parental chromosomes to form those for the child. The blue chromosome is that taken from the father, and the white is taken from the mother.

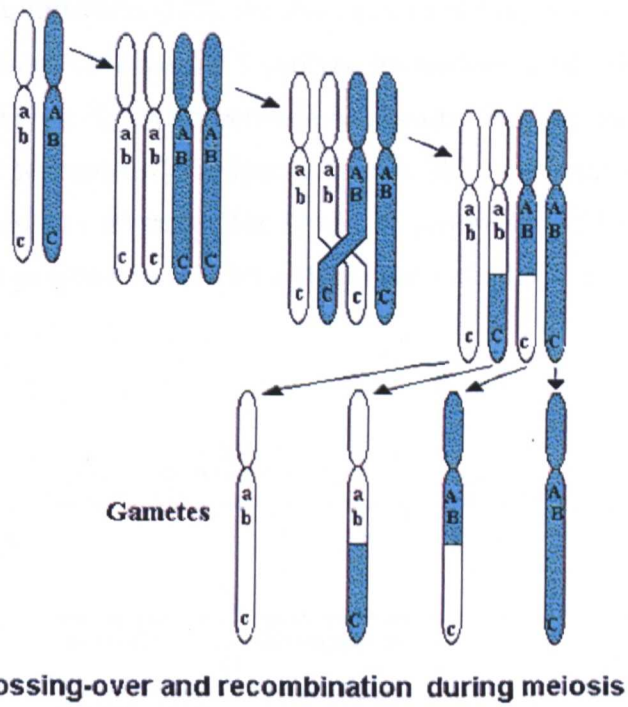


Figure 2.5: Recombination of Chromosomes [5]

It is possible for more than one cross over to occur between two loci.

The recombination fraction is the proportion of offspring that receive a recombinant haplotype from their parents. Usual genotyping methods cannot determine haplotypes directly since genotyping only reports the unordered alleles at each locus and does not report which strand each allele belongs to. If an individual has genotypes A_1/A_2 and B_1/B_2 it is not possible to determine whether the haplotypes present are A_1-B_1 and A_2-B_2 or A_1-B_2 and A_2-B_1. If, however, we have a sample of individuals, it is possible

to estimate the frequencies of each haplotype based on their genotypes and the sample haplotype frequencies. The haplotypes can be imputed using haplotype phasing techniques as discussed in Section 2.7.3.

2.5 Linkage Disequilibrium

When alleles at different but nearby loci are statistically associated they are said to be in **linkage disequilibrium** (LD). Another way of putting this is that LD exists if there is departure from the expected haplotype frequencies if the loci were inherited independently. LD typically exists between two nearby loci that have been inherited together over many generations [6]. Figure 2.6 below shows a stretch of DNA inherited from common ancestors over time. The blocks of yellow are alleles in LD inherited together over many generations. The blue blocks are new alleles introduced by recombination.

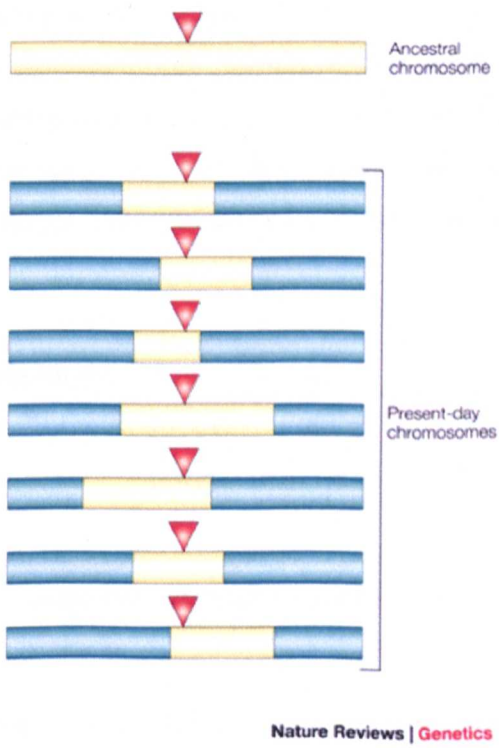


Figure 2.6: Linkage Disequilibrium [7]

Whilst the association between two alleles located next to each other will break down over time due to recombination; it may be maintained over many generations if recombination is low. The two main measures of LD are based on the statistic D. Suppose we have two loci with major alleles A_1 and B_1, and minor alleles A_2 and B_2. A minor allele is one which is the least common for a particular genotype in a given population whereas a major allele is the most common one. A_1 has population frequency p, and B_1 has population frequency q. Assuming no LD, and alleles occurring independently, the expected haplotype frequencies are shown in the table below.

Haplotype	A_1	A_2	
B_1	pq	q(1-p)	q
B_2	p(1-q)	(1-p)(1-q)	1-q
	p	1-p	1

Table 2.3: Haplotype Frequencies assuming no LD (i.e. independence of inheritance)

Let θ be the observed haplotype frequency of A_1B_1. D (a measure of LD), is defined by $D=\theta-pq=0$. D is a measure of LD which is defined as the departure from the frequencies under independent inheritance H_0. We expect that θ = pq when both alleles are inherited independently so that D=0. If LD exists then the haplotype frequencies in the above table no longer hold true and D can lie between -1 and 1.

Haplotype	A_1	A_2	
B_1	pq + D	q(1-p) - D	q
B_2	p(1-q) - D	(1-p)(1-q) + D	1-q
	p	1-p	

Table 2.4: Haplotype Frequencies with LD

Given observed data of p and q and haplotype frequencies, it is simple to estimate D, the measure of LD.

D' and r^2 are the usual measures of LD. They are both diallelic measures and are both used in this PhD thesis.

D depends on the frequency of alleles. D' is a normalised measure, achieved by dividing D by the theoretical maximum given the observed allele frequencies.

$$D' = \frac{D}{D_{max}} \tag{2.1}$$

where

$$D_{max} = \begin{cases} min(p(1-q),(1-p)q) & if D > 0 \\ min(pq,(1-p)(1-q) & \text{otherwise} \end{cases}$$

r is defined by

$$r = \frac{D}{\sqrt{p(1-p)q(1-q)}} \tag{2.2}$$

r^2 is the correlation coefficient and is a measure of similarity between two markers with respect to their minor allele frequencies (MAFs). A measure of r^2 = 1 or D' = 1 between two loci represents "complete" dependency, whereas r^2=0 or D'=0 between two loci indicates independence.

Although the values of r^2 being 0 or 1 and D' being 0 or 1 can be interpreted the same way in terms of independence, the relationship between the two measures of LD is not that simple. D' = 1 when one of the four possible haplotypes is not observed indicating there has been novel genetic variation but not recombination. When D'=1 this does not imply that r^2 will also be 1. They are only equal when both 0-0 and 1-1 haplotypes do not occur, or both 0-1 and 1-0 do not occur. D' is usually used to measure the extent of recombination between loci over several generations, whereas r^2 is usually used to measure similarities between loci, and quantify how well one locus can predict the value of another.

Genetic association studies often have the problem that many genotypes in LD are found to be associated with the outcome of interest. This makes it difficult to establish the actual SNP responsible. This problem will be discussed more in chapter 5.

2.6 SNPs

Single nucleotide polymorphisms (SNPs) occur when a single nucleotide (A,T,C or G) varies between individuals at the same marker. e.g. instead of an A allele there is sometimes a C. The minor allele is that which has the lowest frequency at a locus, usually of two alleles (biallelic) [1]. Figure 2.7 below shows an example of a SNP between two chromosomes.

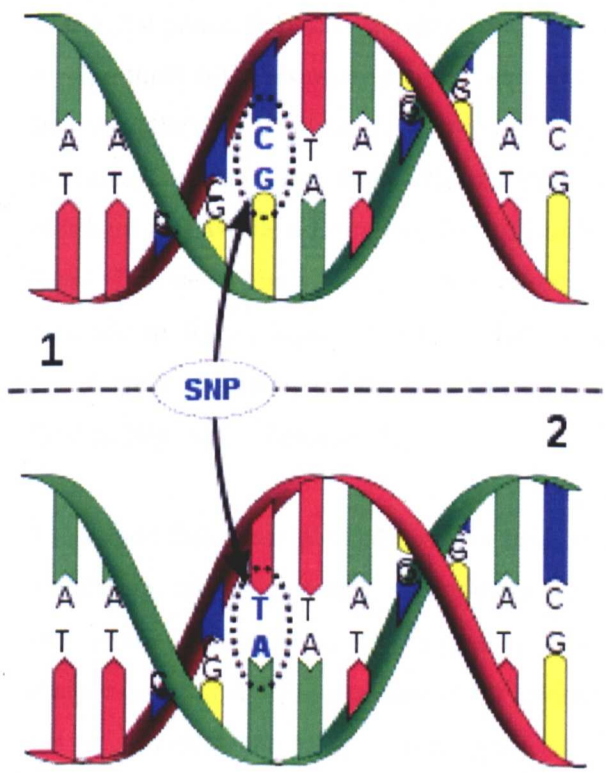


Figure 2.7: Example of a SNP

2.7 Uses of LD

2.7.1 International HapMap Project

The International HapMap project [8] started in 2002 and aims to better understand LD/ haplotype structures. By analysing the patterns of variation in the genome, it is hoped that this information could be used to identify genes associated with complex diseases. "Genetic data from more that one population will enhance the ability of researchers to study the genetic contributions to diseases that are more or less prevalent in different groups."

The HapMap project has collected DNA from 270 people from 4 main populations: U.S. residents with northern and western European ancestry; Yoruba residents from Ibadan, Nigeria; Japanese individuals from Tokyo and Chinese individuals from Beijing. The DNA was used to type approximately 1 million SNPs. Due to the success of the first phase, the study was extended. The second phase used these blood samples to type three times as many markers and was published in 2007. [9] The third phase increased the number of samples from 270 to 1,301 and includes a wider variety of populations. These are those with African ancestry in Southwest USA; Utah residents with Northern and Western European ancestry(as before); Han Chinese in Beijing (as before); Chinese in Metropolitan Denver, Colorado; Gujarati Indians in Houston, Texas; Japanese in Tokyo, Japan (as before); Luhya in Webuye, Kenya; Mexican ancestry in Los Angeles, California; Maasai in Kinyawa, Kenya; Toscani in Italia; and Yoruba in Ibadan, Nigeria (as before). [10]

It is hoped that these populations will help to identify the most common haplotypes worldwide, and help in analysing variation between them. HapMap can be used to view genome information e.g. in haplotype blocks of LD. From this, it is possible to analyse which loci represent most of the underlying variation of the genome. HapMap can act as a reference panel for imputation (detailed under 2.7.3) of loci that may have been untyped in a particular study based on unobserved genotypes at nearby loci.

HapMap has helped us better understand the underlying genetic structures across populations including LD structure.

2.7.2 Tagging

Regions of high LD between loci means that we only need to type a small subset of this region to represent it genetically. We can gain the same amount of genetic information by genotyping a smaller subset. The aim of 'tagging' SNPs is to select those that will best represent the genome or region of interest. [11–14]

2.7.3 Imputation

Genotypes at a particular locus may be missing for several reasons. They may have simply not been typed for the study in question, there may have been a genotyping error, or maybe a reason specific to that allele/locus. Missing genotypes can lead to loss of power and sometimes this missing data can lead to incorrect results due to bias [15].

Imputing missing information leads to a higher mapping density of data. Imputation can increase the power of genetic association. [16]. Imputation can also aid in fine mapping since denser loci can be imputed based on those tagged to be representative of a region of LD. Each locus can then be used for association analysis whether or not they have been typed in the study in question or imputed.

It is common that different studies will type different SNPs across the region of interest. Pooling results can give a better picture and add more power to analyse which variants within a region have the largest effect. If not all loci are typed in each study, imputation can help fill in the missing data. It has been shown that studies with weak findings can be combined together using these methods to identify completely new and highly significant variants.

In terms of imputation, a cluster is a group of loci in high LD. These clusters vary by size along the chromosome.

The program used in this PhD for imputation, Mach [17], makes use of a Hidden

Markov Model (HMM) in updating the EM algorithm to find a maximum posterior distribution. This model states that the unobserved cluster locus depends only on the (unobserved) cluster which the previous locus belongs to.

Mach is one of the most commonly used imputation programs used for large datasets such as GWAS [18]. Mach has been shown to perform as well as other programmes also suitable for large datasets in terms of accuracy and computing time. [19–22]

Mach has the option of calculating the average allele dosage score for each SNP imputed. There is also the option of estimating LD of the imputed SNP with non missing loci.

2.7.4 Rare Variants

It has historically been difficult to detect rare variants due to the tagging of representative SNPs in LD. These tagged SNPs do not represent rarer SNPs well. Whole genome sequencing will hopefully provide more information on these. For example, the 1000 Genomes Project [23]. These types of genotyping collection are possible now the cost of genotyping has significantly reduced in recent years. It is hoped that the information gained by analysing rare variants will help us to better understand complex diseases as we will have a clearer picture of the whole genome.

2.8 Genetic Association Studies

Genetic association studies can be used to determine whether there is an association between a genotypic variant of interest and a disease or trait. For a binary disease outcome (eg. disease/ no disease) and a case control study design, this is done by comparing genotype or haplotype frequencies at the locus of interest by outcome group. Under the null hypothesis of no association, the genotype frequencies between cases and controls should be equal. A case-control study compares a risk factor across two groups; one with disease (cases) and one without (controls). In association studies the

risk factor is the genotype variant [6]. A simple test for association is, for example, to use logistic regression of case status (y) on coded genotype classes:

$$y_i \sim Bern(\pi_i) \tag{2.3}$$

$$\text{logit}(\pi_i) = \alpha + \beta x_i \tag{2.4}$$

where y_i defines the case status for individual i (0 for control, 1 for case), π_i is the probability of individual i being a case, α is the intercept, x_i is the genotype variant of subject i; usually coded 0 for first homozygous genotype, A/A for example, 1 for the heterozygous genotype, A/T for example, and 2 for the other homozygous genotype, T/T for example, and β is the genotypic effect on probability of disease. This is an **additive** model on the log scale as the odds ratio (OR) for T/T is twice that of A/T i.e. each T allele increases the OR by an equal amount. Other models can be considered to allow for a dominant or recessive effect of the genotype, as described below.

Under a **dominant** model, only one copy of the variant allele is required to cause an increase in risk of the disease. Having two copies of the variant allele is assumed not to increase that risk. i.e. $P(D|AA) = P(D|Aa)$.

$$\text{logit}(\pi_i) = \alpha + \beta I(x_i \neq 0) \tag{2.5}$$

Under a **recessive** model, both copies of the variant allele are required to cause an increase in risk of the disease. $P(D|aa) = P(D|Aa)$

$$\text{logit}(\pi_i) = \alpha + \beta I(x_i = 2) \tag{2.6}$$

Under a 2 degrees of freedom (2d.f.) model, the increase of risk of disease by having two copies of the variant allele is different to that of having only one copy, but not in an additive way.

$$\text{logit}(\pi_i) = \alpha + \beta I(x_i = 1) + \gamma I(x_i = 2) \tag{2.7}$$

The appropriate test of association now has 2d.f. and tests against the null hypothesis which in this case is $\beta = 0$ & $\gamma = 0$. β can be viewed as the additive genetic component and γ as a dominance (or equally recessive) component or the deviation away from the additive model.

It is possible to show the genotype data for each SNP in a simple contingency table, as shown below.

Table 2.5: Contingency Table for One SNP

Genotype	Cases		Controls		Total
	Observed	Expected	Observed	Expected	
0	n_{0Ca}	$\frac{n_{0.}n_{.Ca}}{n_{..}}$	n_{0Con}	$\frac{n_{0.}n_{.Con}}{n_{..}}$	$n_{0.}$
1	n_{1Ca}	$\frac{n_{1.}n_{.Ca}}{n_{..}}$	n_{1Con}	$\frac{n_{1.}n_{.Con}}{n_{..}}$	$n_{1.}$
2	n_{2Ca}	$\frac{n_{2.}n_{.Ca}}{n_{..}}$	n_{2Con}	$\frac{n_{2.}n_{.Con}}{n_{..}}$	$n_{2.}$
Total	$n_{.Ca}$	$\frac{n_{..}n_{.Ca}}{n_{..}}$	$n_{.Con}$	$\frac{n_{..}n_{.Con}}{n_{..}}$	$n_{..}$

Under the null hypothesis of no association with the disease, it is expected that the genotype frequencies are the same in cases and controls. A 2d.f. score test for association can be calculated using Pearson's χ^2 statistic for independence of the rows and columns given by

$$\chi^2_{Gen} = \sum_{i=0,1,2} \sum_{j=Ca,Con} \frac{(n_{ij} - \mathrm{E}[\mathrm{n_{ij}}])^2}{\mathrm{E}[\mathrm{n_{ij}}]} \tag{2.8}$$

$$\mathrm{E}[\mathrm{n_{ij}}] = \frac{\mathrm{n_{i.}n_{.j}}}{\mathrm{n_{..}}} \tag{2.9}$$

Under the null hypothesis of no association (independence between the rows and columns of the table above) of SNP and outcome, the test statistic has an approximate χ^2 distribution with two degrees of freedom.

2.8.1 Fisher Exact Test

An alternative to the Pearson's χ^2 test is the Fisher exact test which avoids relying on asymptotics. This is especially useful when SNPs have small frequencies of genotypes. In addition, a Fisher exact test does not assume an additive model, and allows for any type of association between the SNPs and outcome. It may not be that all SNP associations are additive, and this test gives more flexibility by SNP. In this framework, each SNP is tested for an association individually. The Bayesian Multivariate Adaptive Regression Spline model I use later (5.9) has a dominance component for such flexibility, and using a Fisher exact test for my frequentist analysis allows for more direct comparisons to be made under the different methods.

As explained above, genotypes can be coded as 0, 1 or 2 (two copies of the minor allele, one copy of each the minor and major allele, or two copies of the major allele). It is possible to show the genotype data for each SNP in a simple contingency table, as shown above.

Under a Fisher exact test [24] the probability of obtaining the observed values in the table above is given by the hypergeometric distribution. This probability is given by

$$p = \frac{\binom{n_{0.}}{n_{0Ca}}\binom{n_{1.}}{n_{1Ca}}\binom{n_{2.}}{n_{2Ca}}}{\binom{n_{..}}{n_{.Ca}}} \tag{2.10}$$

2.8.2 Data Augmentation in Binary Probit Regression

In order to simplify modelling a binary outcome, a probit link function with data augmentation can be used.

By introduction of latent variables (via data augmentation) it is possible to reduce a test of association with a binary outcome to a Gaussian linear model [25]. Consider

$$y_i \sim Bernoulli(\Phi(\eta_i)) \tag{2.11}$$

where

$$\eta = \alpha_1 + \sum_{k=1}^{K} \beta_k \mathbf{x} \tag{2.12}$$

In the probit model, the mean is given by $\mu_i = \Phi(\eta_i)$ so it follows that the probit link function $g(\mu_i) = \Phi^{-1}(\eta_i)$

Introducing a set of latent variables w_i for the i^{th} observation with a Gaussian distribution conditional on observation specific random terms.

$$w_i \sim N(\eta_i, 1) \tag{2.13}$$

such that

$$y_i = \begin{cases} 1 & \text{if } w_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the distribution of y_i having integrated out w_i is

$$
\begin{aligned}
P(y_i = 1) &= P(w_i > 0) = P(N(\eta_i, 1) > 0) \\
&= P(N(0,1) > -\eta_i) = P(N(0,1) < \eta_i) = \Phi(\eta_i)
\end{aligned}
\tag{2.14}
$$

2.9 Potential Problems with Genetic Association Studies

As explained in Section 2.3, genotypes are passed on to offspring under Mendelian randomisation. This means that genotypes are inherited at random with equal probability. Genotypes are invariant to mRNA, proteins, diseases and environmental factors. Therefore, genetic associations should be protected from reverse causation as these things cannot cause a particular genotype. In addition, environmental factors cannot be considered as possible confounders in a genetic association test which is a possible problem with other association tests.

However, genetic association studies can still suffer from selection bias. For example, in population based association studies, it is necessary to ensure that the cases and controls have the same ethnic background otherwise a gene that differs between ethnic groups could appear to be associated with the disease if disease prevalence differs in the two populations.

Another potential problem is that two loci could be so closely in LD with each other that they both appear to be equally associated with the disease outcome. In this case it can be hard to determine which locus is truly associated with the disease by considering just the single SNP association tests. More complicated approaches that try to correct for the effect of other loci are required to understand the data more clearly.

Finally, there is the problem about which disease model to assume. Often the additive model is used but this can lead to loss of power when this model is incorrect, especially in the recessive case. Methods that do not force a particular model are desirable.

CHAPTER

3

BACKGROUND TO STATISTICAL THEORY

The majority of this Chapter is referenced from Bayesian Data Analysis [26], Markov Chain Monte Carlo in Practice [27] and Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference [28] unless otherwise stated.

3.1 Statistical Inference in a Frequentist Setting

In classical or frequentist statistics, observed data X=x are used to make inferences about a population parameter θ which we consider to be **fixed,** i.e. true but unobserved. One approach to estimate θ in a frequentist setting is via likelihood modelling. Suppose that $X_1, ..., X_n$ are observable random variables with a joint distribution that depends on unknown parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$. The likelihood function of $\boldsymbol{\theta}$ is found by evaluating this distribution at the observed data (sample) $\boldsymbol{x}=(x_1, ..., x_n)$, $f(\boldsymbol{x}|\boldsymbol{\theta})$. Note: This is not a probability distribution for $\boldsymbol{\theta}$ as it does not sum to 1 over θ.

As the data x are known, we are only interested in how the likelihood varies with θ. Arguably the best estimator of the true value of θ is that value of θ which maximises the likelihood function. This estimator, $\widehat{\theta}$, is known as the maximum likelihood estimator. [29–32]

Inference can take the form of a **point estimate** (for example, $\widehat{\theta}$=0.1); a **confidence interval** (range in which θ will lie within with a specified probability); a **hypothesis test** (for example, reject the hypothesis that $\theta < 0.07$ at the 5% significance level); a **prediction** (predict that 15% of patients will have an adverse event); or a **decision** (decide to stop treatment on patients with adverse events). In each case, knowledge of the observed sample value X=x is being used to draw inferences about the population characteristic θ. Moreover, those inferences are made using the likelihood function, $f(x|\theta)$, which determines how, for a given value of θ, the probabilities of the different values of X are distributed. In this setting of frequentist statistics, the statistical parameter, θ, although it is unknown, is treated as a **constant** to be estimated rather than as a random variable.

3.2 Bayesian Inference

Bayesian inference allows us to combine the knowledge from observed data, and any prior knowledge we may have before the sample is collected. It also allows us to make inference about the distribution of the parameter values.

The fundamental difference between frequentist and Bayesian statistics is that in a Bayesian context, θ is treated as a **random** (vector) variable. Before collecting data on the random variable **X** (which possibly depends on θ), the distribution of θ is believed to have prior density f(θ). The probability distribution of θ is updated given the observed data ($f(x|\theta)$), using Bayes theorem, to give the posterior distribution $f(\theta|x)$, which is the probability of the parameter θ given the observed data x. Inference is based on the posterior, rather than the likelihood.

Bayesian statistics revolves around Bayes theorem, which defines the posterior distribution as:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} = \frac{f(\theta)f(x|\theta)}{\int(f(\theta)f(x|\theta)d\theta} \propto f(\theta)f(x|\theta) \text{ [1]}$$

where f(θ) is the prior density, encompassing our prior beliefs about θ, and f($x|\theta$) is the likelihood; the same as that used in frequentist inference outlined above. $\int(f(\theta)f(x|\theta)d\theta$ is the normalising constant used to make f($\theta|x$) a probability density (i.e. sum to 1). This normalising constant often does not need to be defined explicitly since $f(\theta)f(x|\theta)$ includes all the information about θ; the random vector we are interested in making inferences about.

The prior distribution, f($\boldsymbol{\theta}$), which represents our prior knowledge or beliefs about $\boldsymbol{\theta}$ could be, for example, obtained from results of previous studies, or from expert opinion.

Consider a simple example. If we have three experiments:

1. A tea-drinker claims she can tell whether the milk was added before or after the tea. Out of 10 attempts, she is correct 10 times.
2. A music expert claims she can identify and name any piece of Mozart's work. She correctly identifies 10 out of 10 pieces.
3. A drunk friend claims she can predict the outcome of the toss of a fair coin, and does so 10 times in a row.

In all of the above, the model is $f(X|\theta) \sim Bin(10, \theta)$ and x=10 is observed.

In frequentist statistics, using $f(x|\theta)$, we would make the same inferences about θ in each case. Opinions differ as to whether this is either a draw back or an advantage of inference in a frequentist setting. Our prior beliefs are likely to be different in each of the above situations. Our prior beliefs are likely to remain highly skeptical about 3,

partly convinced by 1, and perhaps not surprised by 2.

It can sometimes be difficult to express prior beliefs, and an uninformative prior might also be used in this case.

3.3 Inference

If a posterior distribution has the same parametric form as the likelihood distribution, for the parameters of interest, then the prior is known as a conjugate prior. Conjugate priors are often convenient. For example, if the prior distribution is Gaussian, and the observed data has a likelihood distribution that is also Gaussian with known variance, then the posterior distribution is also Gaussian. In this case, there is no need to calculate the constant of proportionality of the full posterior distribution because this is already known from the distributional form. [26]

Suppose we have data y=$(y_1, \ldots, y_n)$ that are i.i.d. with Gaussian distribution with likelihood

$$p(y|\theta) = \prod_i^n p(y_i|\theta) \propto \prod_i^n exp(-\frac{1}{2\sigma^2}(y_i - \theta)^2) \tag{3.1}$$

where the variance σ^2 is assumed to be fixed at some known value. With a Gaussian prior with mean μ_0 and variance τ_0^2, the posterior is also Gaussian:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &= p(\theta)\prod_i^n p(y_i|\theta) \\ &\propto exp(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2)\prod_i^n exp(-\frac{1}{2\sigma^2}(y_i - \theta)^2) \\ &\propto exp(-\frac{1}{2}[\frac{1}{\tau_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\sum_i^n (y_i - \theta)^2]) \end{aligned} \tag{3.2}$$

leading to

$$p(\theta|y) = N(\theta|\mu_1, \tau_1^2) \tag{3.3}$$

with $\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} and \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$

However, it is not always possible to find a conjugate prior distribution that fits our beliefs, or perhaps a conjugate distribution is not known for the likelihood of interest. In non conjugate settings the implementation of Bayes Theorem can be computationally difficult, usually as a result of having to calculate the normalising integral in the denominator. Calculating this normalising integral is necessary if we want to make most inferences or predictions from the posterior distribution. Calculating this integral can be extremely computationally intensive. As mentioned above, for some choices of the prior distribution, calculating this integral can be avoided, but in general, specialised techniques are required to perform this calculation. i.e. Inference from the posterior distribution can be done either algebraically, or computationally using sampling. One such method of sampling is Markov Chain Monte Carlo (MCMC).

3.4 Markov Chain Monte Carlo

Markov chain simulation algorithms allow us to sample from the posterior distribution when calculating its full distribution is algebraically difficult. i.e. when the resulting posterior distribution is non-tractable. MCMC methods simulate a Markov chain, whose stationary distribution is the posterior distribution we are interested in. [28,33]

A Markov chain is a sequence of random variables Z_1,Z_2,Z_3,... with the property that at each time point t, the next state Z_{t+1} is sampled from a distribution dependent only on the current state Z_t. The possible values of Z_t form the state space, S. This is called the Markov property: given the present state, the future and past states are independent.

Under certain conditions (see below), the Markov chain converges to a **stationary** distribution that is, one which does not change over time. This stationary distribution does not depend on t or Z_0.

A chain converges to a stationary distribution if it satisfies all three of following conditions [28,34]:

1. It is irreducible: there is a probability that the chain can assign any possible member of state space S to Z_t in a finite number of iterations.
2. It is aperiodic: the chain does not cycle between a subset of values for Z_t in a regular periodic movement.
3. It is positive recurrent: given any initial value of Z_t, the expected number of iterations to return to that initial value is finite.

Note: Each of these conditions by themselves is necessary.

In MCMC, we simulate from the "target" (posterior) distribution, making enough draws so the distribution of draws is hopefully "close enough" to the stationary distribution. Once the stationary distribution has been achieved, future draws from this distribution are still dependent since every new draw is now sampled conditionally on the previous state.

When simulating using MCMC algorithms it is required to check convergence with plots, summaries, and then delete the first M simulated values as a burn-in period. The burn-in period is the time which it takes the algorithm to reach a stationary distribution. Once we are satisfied the data has converged, the target (posterior) distribution can be summarised by the simulated values of θ drawn after this point. For example, an approximation to the mean of the distribution can be found simply by taking the mean of the simulated values of θ.

3.5 Metropolis Hastings Algorithm

One such example of a MCMC algorithm is the Metropolis Hastings (M-H) algorithm. Suppose we have a vector θ of dimension d of parameters we wish to sample. The basic outline of a M-H algorithm is to

1. Select a starting point for vector θ
2. Propose a candidate value of θ for the next step of the Markov Chain
3. Accept the proposed value of θ given the rule below
4. Repeat the steps iteratively

In general terms, the chain is initialised with $\theta_1^0,\ldots,\theta_d^0$. Now suppose the current values of the chain is $\theta_1^j,\ldots,\theta_d^j$ and that we want to simulate θ_1^{j+1}, the next value of θ_1. The general scheme of the MCMC is to update θ_1^j to θ_1^{j+1} and accept the new value using the acceptance rule below.

Schematically the general Metropolis-Hastings updating mechanism is:

- Propose a candidate value θ_1^{can}, which is a drawn from an arbitrary distribution with density $q(\theta_1^{can}|\theta_1^j,\theta_2^j,\ldots,\theta_d^j)$.
- Take as the next value of θ_1 in the chain

$$\theta_1^{j+1} = \begin{cases} \theta_1^{can} & \text{with probability p} \\ \theta_1^j & \text{with probability 1-p} \end{cases}$$

where

$$\text{p} = \min\left(1, \frac{p(\theta_1^{can},\theta_2^j,\ldots,\theta_d^j|x)}{p(\theta_1^j,\theta_2^j,\ldots,\theta_d^j|x))} \frac{q(\theta_1^j|\theta_1^{can},\theta_2^j,\ldots,\theta_d^j)}{q(\theta_1^{can}|\theta_1^j,\theta_2^j,\ldots,\theta_d^j)}\right)$$

with $p(\theta_1^{can},\theta_2^j,\ldots,\theta_d^j|x)$ denoting the posterior distribution of θ_1 evaluated at $\theta_1 = \theta_1^{can}$ and similarly for $p(\theta_1^j,\theta_2^j,\ldots,\theta_d^j|x)$ evaluated at $\theta_1 = \theta_1^j$.

- Update each member of θ, proposing θ_2^{can}, θ_3^{can}, etc. in this way
- Iterate this procedure

The candidate generator $q(\theta_1^{can}|\theta_1^j, \theta_2^j, \ldots, \theta_d^j)$ is arbitrary but some choices of q(.) will lead to faster convergence, and are therefore more computationally efficient. The variance of the candidate distribution is an important choice because if it is too big then the proposed moves will be too large, and acceptance probabilities will be low. However, if variance is chosen to be too small then then the acceptance probabilities will be high but only small steps will be taken, and convergence will be slow.

Note: a common choice for the candidate generator is the density of a Gaussian distribution for θ_1^{can} with mean θ_1^j. This is known as the Random Walk Metropolis Hastings algorithm. Due to the symmetry of this candidate generator, the terms in the acceptance probability involving q(.) cancel and this reduces to

p= min(1, ratio of posterior distribution of θ_1^{can} vs θ_1^j)

It can be shown that the Metropolis-Hastings algorithm converges to a stationary distribution, equal to the target posterior distribution.

3.6 Bayes Factors

Bayes Factors (BF) are increasingly used in genetic epidemiology as an alternative to frequentist p-values. If we have a discrete set of possible models, a Bayes Factor is the ratio of posterior to prior odds of one model compared to another. If we wish to compare two models M_i and M_j the Bayes Factor is defined as

$$BF(M_i, M_j) = \frac{p(M_i|D)}{p(M_j|D)} \frac{p(M_i)}{p(M_j)} \tag{3.4}$$

where D is the data given. If the two models have equal prior probabilities then $p(M_i)/p(M_j) = 1$ and BF(M_i, M_j) is simply $\frac{p(M_i|D)}{p(M_j|D)}$ [26] , [35], [36]

For example, a Bayes Factor can be used to test one model against the null hypothesis in a linear regression that $\beta = 0$. BFs > 3 are usually interpreted as an indication of evidence in favour of M_i. [37]

3.7 Reversible Jump Metropolis Hastings Algorithm

Sometimes we wish to sample over models with varying dimensions. For example, with regression models, it is likely that we will want to select the most important predictors among a sometimes large set of variables. As in genetic applications, for example.

One solution to this is to sample over the model space, and treat the model structure (which variables, and how many) as an additional, separate parameter, say ξ. We are then interested in the posterior distribution of this parameter. For example, consider a simple regression model with 10 possible explanatory variables. If we were to propose the first, third and eighth variables in the model, the parameter space for this iteration could be defined by ξ=(1,0,1,0,0,0,0,1,0,0) where 1 indicates the inclusion of the corresponding explanatory variables (say, β).

The Reversible Jump MCMC [34, 38] scheme deals with this. The Reversible Jump algorithm is an extension of the Metropolis-Hastings algorithm, and samples from posteriors of varying dimension. At each step of the algorithm, we propose to either add a variable to the current model (increase the dimension of corresponding explanatory variables β by 1) which is known as a 'birth' step, or drop one (decrease the dimension by 1) which is known as a 'death' step. Note: At each iteration, a new value is proposed for the element of β relating to the variable in question. i.e. a new value is randomly proposed when adding a term, but is simply forced to 0 if dropping the term.

These steps are chosen at random, and the proposed vector of parameter space (ξ') is accepted with probability

min (1, likelihood x prior x proposal ratio)

For example, if a birth step is proposed the acceptance probability is

$$\min(1, \mathrm{BF}(\xi', \xi)\frac{d_{k+1}}{b_k})$$

where d_k is the probability of dropping one element of ξ and b_k is the probability of adding one element of ξ. If $b_k = d_k = 0.5$ then the acceptance probability is simply $\min(1, \mathrm{BF}(\xi', \xi))$.

Note: The above acceptance probability also includes a Jacobian term to account for the change in dimension between a model with parameter space ξ and ξ' but in practice this is rarely needed. [34]
In this situation, the MCMC algorithm is set up to combine a Reversible Jump algorithm to move in model space, ξ (ξ' is accepted with above probability), and then a M-H sampler as described in Section 3.5 to draw values of the current corresponding explanatory parameters in the model, β.

By monitoring both ξ and β, this algorithm would give posterior probabilities of the models visited as well as the usual posterior distribution of model parameters. It is then possible to decide the importance of each predictor by summing the posterior probabilities of the models containing the relevant term. This gives the marginal probabilities of each predictor. It is also possible to examine the joint probabilities of variables.

MCMC schemes rely on being able to sample parameters conditional on the value of others. Directed acyclic graphs (DAGs) represent such dependencies naturally and conditional independence structures can be represented graphically. We discuss this in the next chapter.

CHAPTER

4

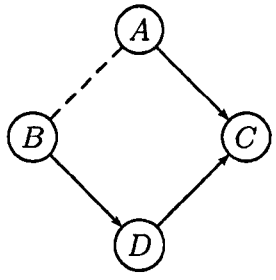
INTRODUCTION TO DIRECTED ACYCLIC GRAPHS

4.1 Graphical Models

Graphical models are one way to present statistical relationships. Precisely, dependencies between variables and indicating conditional independent structures. Using them it is possible to represent assumptions about relationships between variables. Fitting graphical models also helps to determine whether is it possible to identify directions of association with the data available and they can highlight possible biases. In addition, graphical models make it easy for the reader to understand or picture more complex relationships and can help set up joint probability models for such complex data systems.

A graphical model has **nodes** representing variables. Any line or arrow connecting two variables in a graph is called an **edge.** Edges can be **directed** (represented by a single-headed arrow) to represent direct links from one variable to another; or non-directed (usually represented by a dashed line). Edges represent direct associations between

variables after accounting for all other variables in the graph. In the graphical model below, for example,

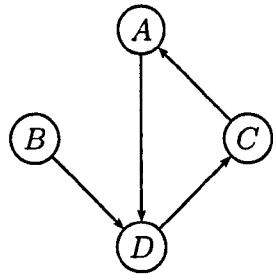


B has a direct effect on D, D has a direct effect on C, A has a direct effect on C and A and B are associated but direction of effect is not specified. The association between B and C is entirely through A and D. i.e. indirect. [39]

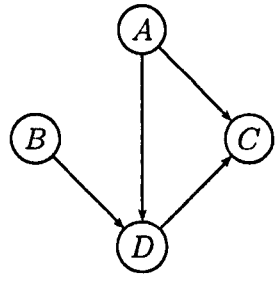
A **path** in a graphical model is defined as a sequence of edges connecting one variable to another. A path can have directed or undirected edges, and need not follow the direction of the edges.

4.2 Directed Acyclic Graphs

A directed acyclic graph (DAG) is a graphical model with directed edges, and no closed loops (i.e. for all variables in the graph there does not exist a directed path from a particular variable to itself). For example, the graph below contains a cycle (A to D to C to A), and is therefore **not** a DAG. [40]



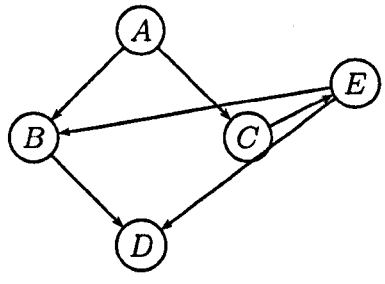
On the other hand



is a DAG.

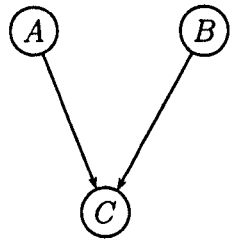
4.3 DAG Terminology

Consider this simple example of a DAG:



Some simple terminology of DAGs, using the above diagram as an example: [39,40]

- Two variables are **adjacent** if they are directly connected by an edge. e.g. A and B are adjacent, but A and D are not.
- A **path** through the graph is any unbroken route connecting adjacent nodes.
- A **directed path** through the graph is any unbroken route connecting adjacent nodes by directed edges and following the direction of these edges e.g. A to B to D
- A variable is known as an **ancestor** if there is a directed edge or path from that variable to another. e.g. A is ancestor of B,D,C and E.
- A variable is known as a **parent** of another if there is a directed edge from it to another adjacent. e.g. A is a parent to B and C. B and C are also said to be directly affected by A.
- A variable is known as a **descendant,** or affected by another if there is a directed path into that variable. e.g. D is a descendant of C.
- A variable is known as a **child** if there is a directed edge into that variable from another. e.g. D is a child of B and E.
- A **backdoor path** is defined as one in which there the first variable in the path is a child of the second, and there are 3 or more variables connected in the pathway e.g. D to E to B.
- A **collider** is a node with at least two parents. e.g. D or B.
- A path is **blocked** or **closed** if it has one or more colliders on it. e.g. A to B to D to E.
- A path is **unblocked** or **open** if there are no colliders on it. e.g. D to E to C is an unblocked path. In this case it is also a backdoor path.
- A **v-structure** is a collider in which the parents of the colliding node are not adjacent. The two parents have directed edges towards the same child, creating a 'v'. e.g. the following DAG is a v-structure:



In this given v-structure, A and B may be marginally independent as they do not have an association between them directly, but they are dependent conditionally on C as there is a pathway from A to B through C.

4.4 Using DAGs in Statistical Modelling

Quantitative statistical approaches can be used to translate DAGs into statistical models. For example, statistical models can be represented using DAGs, showing the joint relationships between variables. This representation can be used efficiently for defining joint probability distributions, and possibly drawing conclusions about direct associations. DAGs can be used to encode conditional independent structures, and generate convenient factorisations of a joint distribution.

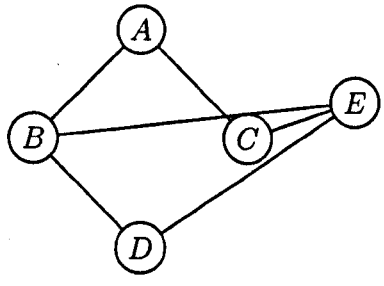
Our naive hope in using DAGs for inference is that directed edges within our DAG may help to suggest directions of associations. In fact, directions of association can be very difficult to infer from such DAGs for two main reasons. Firstly, it is very difficult to be sure that all unobserved confounders have been accounted for in observational studies and secondly there is the problem of DAG equivalence classes. [39,40]

4.5 Equivalence Classes

Different DAGs can be shown to infer the same underlying 'conditional independence' model. For example,

have the same joint distribution. In this case directions of association become indistinguishable. We are interested in the implications of such limitations. Two DAGs are equivalent if [40]

1. They have the same undirected graph. i.e. same graph but without the direction of association on it (no arrows); only an indication of association between the nodes. For example,

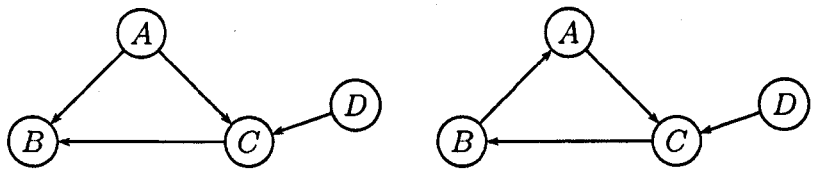


is the undirected graph of the previous DAG shown.

AND

2. They have the same v-structures

Consider another example below.



These DAGs are equivalent because they have the same undirected graph, and the same v-structure A to C and D to C. It is impossible to tell the direction of association between nodes A and B given only this information. Even in a perfect situation with completely observed population based data, equivalence classes can not be distinguished.

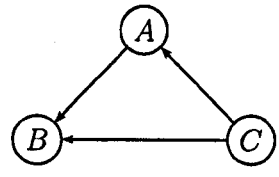
4.6 Association in DAGs

DAGs are non-parametric models in that they do not imply anything about specific distributions between the variables. The directed edges between nodes imply a relationship between variables. If there is no edge then this implies no direct relationship or association.

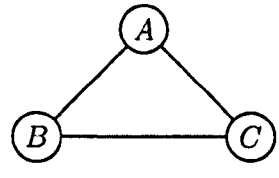
There are two main approaches when trying to establish the true underlying model of a given scenario. Firstly, thinking in terms of directions of association and hypothesising about these between effect and outcome, then testing this model to see if the data supports it, or if it can be falsified. On the other hand, assuming that the model is unknown and trying to use the data to suggest 'likely' models or the most 'likely' model. In this context, one cycles over all possible DAGs to find the one(s) that best fit(s) the data. Given an optimal model or set of models we need to consider the corresponding equivalence classes in order to establish information about directionality. Different equivalence classes may suggest alternative conclusions about the true model. [39]

4.7 Trying to Determine Direction of Association

Essentially, the interest is in trying to analyse which variables have a direct effect on others. I want to allow for everything that could be related to the variables in question, e.g. confounders, to make sure the model is correct. In a statistical framework, DAGs can be used to model all the variables jointly. This will automatically correct for the effects of all variables included in the model via the edges defined. Therefore, I allow an algorithm (described later in Section 6) to choose the most appropriate model. A **confounder** is a variable which is associated with both the outcome and exposure but is not on the causal pathway from exposure to disease. If there is a confounder within the model then this will be automatically corrected for. For example, fitting the model below



will automatically correct for the effect of confounder C assuming primary interest is in effect of A on B. The model search algorithm will decide whether the edge between A and B is necessary when confounder C is taken into account. Note that the DAG shown above is defined by equivalence class



with arrows in any direction since no v-structures are present. If confounder C is not observed, or it is not adjusted for in the model, the true effect between A and B is distorted by the associations of A and B with C.

4.8 Instrumental Variables

Instrumental variables can be used to infer causal relationships.

Consider the following DAG

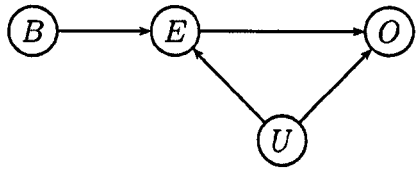


Figure 4.1: DAG Illustrating Instrumental Variable, B

Suppose we are interested in the association, and direction of association between an exposure, E and outcome, O. There may be unobserved confounders influencing this relationship. These are labelled 'U' in Figure 4.1 above.

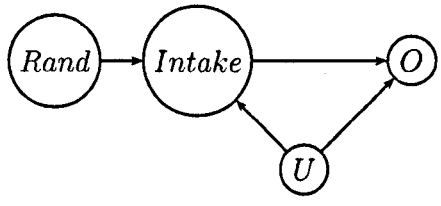
Here, B is an instrumental variable for the relationship from E to O. A variable is defined as an **instrumental variable** if it is

1. associated with exposure, E,
2. has no direct effect on the outcome, O,
3. and does not share common causes with the outcome.

An instrumental variable allows the estimation of the effect of exposure even in the case of unobserved confounders because it is only associated with outcome if exposure is. In other words, B has an association with O but **only** through E. This approach may offer a strategy for eliminating or reducing unobserved confounding, in the estimation of E on O.

The instrumental variable forces the direction between E and O to be known due to the v-structures implied by the added node.

Randomisation of treatment in a randomised control trial, for example, can be considered an instrumental variable, as shown in the DAG below. It is associated with intake of a drug (exposure); has no direct effect on outcome due to blinding; and does not share common causes with outcome due to randomisation.



Genotypes can often be used as instrumental variables in genetic studies. Mendelian randomisation (MR) states that genotypes are assorted randomly at birth with equal probabilities. Genotypes are also not influenced by exogenous (environmental) factors. In this respect, genotype can be used in the same way as a randomised treatment in a clinical trial by potentially allowing an unbiased estimate of the effects of gene products (intermediate phenotype) on outcomes (disease risk/status) i.e. as an instrumental variable. Genotype can be associated with exposure, have no direct effect on outcome and does not share common causes with the outcome. The effects of a genotype on the outcome in this scenario are assumed to be only through the intermediate phenotype.

However, it should be noted that genotypes under the assumptions of instrumental variables through MR have several limitations. This methodology is subject to challenges such linkage disequilibrium, pleiotropy, weak genetic effects and lack of knowledge of how genetic variants biologically effect phenotypes. [41,42]

Under MR it may be possible to analyse the effects of an intermediate phenotype on a disease using genotypes as instrumental variables (IV), in a set up generally free of confounding by environmental exposures. However, confounding by linkage disequilibrium (LD) or population stratification may still occur. Population stratification can be a confounder as different populations carry different risks of disease and genotypes. Depending on the SNPs typed, the analysis of the genotype effect on outcome may be

biased due to a possible omission of untyped disease causing variants in LD with the typed SNPs.

Linkage disequilibrium may also be an issue in that there may be another genotype in LD with the genotype being used as an IV. This could violate the IV assumption that the genotype is only associated with disease/outcome through intermediate phenotype/exposure as shown in the diagram below.

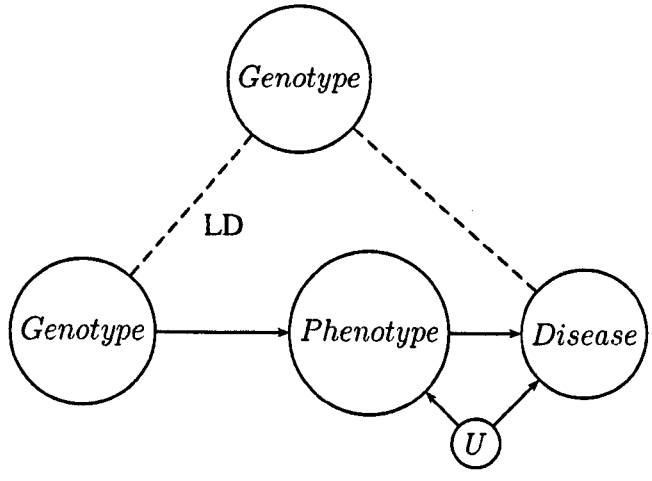


Figure 4.2: Genotype as Instrumental Variable with Possible Genotype in LD

In the same way pleiotropy may be an issue in using a genotype as an IV. If a gene has multiple phenotypic traits; or acts via more than one pathway then the effect of this on the outcome may be confounded by other pathways from gene to outcome. This would, again, invalidate the assumption that genotype is only associated with disease through intermediate phenotype.

The association between the intermediate phenotype and on a disease is usually calculated using a ratio of regression coefficients of association between the variables. If the regression coefficients are as shown in the diagram below

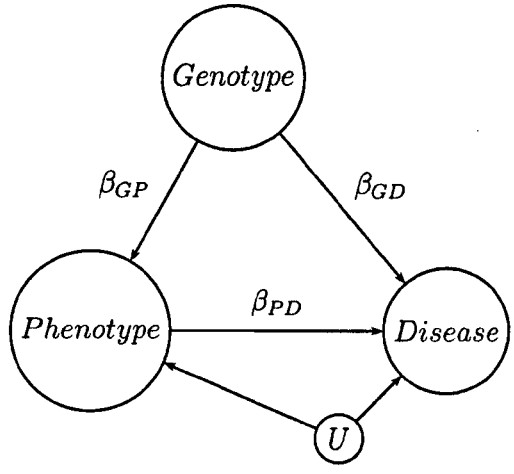


Figure 4.3: Coefficients of Association

then the expected association between the phenotype and disease is given by

$$E(\beta_{PD}) = E(\beta_{GD})/E(\beta_{GP}) \tag{4.1}$$

There will be some degree of bias for E(β_{PD}) when the IV assumptions outlined above are not met.

If the effect of the genotype on the intermediate phenotype is weak then this will lead to uncertainty in the model. Weak instruments can have underestimated confidence intervals. A weak correlation between the instrument and error in the original equation can lead to large inconsistencies. The less precisely the genetic variation predicts the intermediate phenotype, the less precise the derived effect estimate for the association between phenotype and disease will be. The above equation equation requires there to be no large amounts of variation in the numerator. A weak effect can therefore violate the assumptions required of an instrumental variable: an association is required between the IV (genotype) and exposure (intermediate phenotype). Weak instrument bias is in the direction of the confounded association between intermediate phenotype and disease.

There is also the problem of lack of knowledge of how genetic variants have their effects. This may mean that the pathways modelled are missing variables that would

affect the estimates or pathways. Genetic variants may all be interacting and it may not be clear which genotypes to include in the model.

Using genotypes as instrumental variables through Mendelian randomisation should be done with caution, especially if the gene effect on intermediate phenotype is weak.

CHAPTER

5

BAYESIAN MULTIVARIATE ADAPTIVE REGRESSION SPLINE MODELLING

5.1 Aims and Background for SLE dataset analysis

My aim is to investigate Bayesian variable selection methods in regions of high LD. In particular, to investigate SNPs in the major histocompatibility complex (MHC) region associated with systematic lupus erythematosus (SLE). Past studies have found several SNPs in this region to be highly associated with SLE but these SNPs are in high LD.

The major histocompatibility complex (MHC) region on the short arm of chromosome 6 was first fully sequenced in 1999 by the MHC Sequencing Consortium. The gene clusters found to have the most defined functional relevance in terms of antigen processing and presentation were the HLA class I (HLA-A,-B,-C) and class II (HLA

-DP,-DQ,-DR). [43] To date more than 100 diseases, including autoimmune diseases, have been found to be associated with HLA genes, and the MHC region has been found to have SNPs with the highest associations, in most cases, for autoimmune diseases. However, there is a large amount of genetic variation and LD across the MHC which hinders attempts to define the primary signals associated with disease and to determine primary signals. [44] [45]

Rioux et al [46] aimed to investigate the strong linkage disequilibrium across the MHC region by genotyping a very large dataset. They aimed to establish the common genetic variants across the 3.44 Mb region using 10,576 DNA samples. They genotyped 1,472 SNPs, and analysed the genetic associations in this region with several auto-immune diseases including SLE. Systematic lupus erythematosus (SLE) is a disease of the immune system, and can cause inflammation of the joints, and certain organs of the body.

After initial analysis, Rioux et al [46] pooled the UK and US datasets together as the individual SNP analysis for association with SLE gave the same 6 top markers using each dataset. They concluded that their approach was robust, and they had high quality sample collections. With a pooled dataset, the power of their statistical analysis increased.

The analysis showed that the top signal for association with SLE was RS1269852 with an odds ratio of 2.4 and an associated p-value of 5.63E-29. Other top signals were RS558702, RS3130484, RS3131378 and RS3131379 with p-values of 6.75E-29, 1.59E-26, 1.9E-26 and 1.9E-26 respectively; and odds ratios of 2.34, 2.25, 2.24 and 2.24. It was found that these SNPs are all in extremely high LD with RS1269852 with $r^2 > 0.93$.

Conditioning on RS1269852 to find secondary associated SNPs, Rioux et al found RS3135391 to have the highest signal (p-value of 3.9E-06). However, this SNP is also in high LD with RS1269852 (r^2=0.98). They found signals potentially independent of RS1269852 to suggest at least 3 separate signals in this region. The strong LD across the SNPs found to be associated with SLE makes it difficult to identify the causal ones.

The aim of my analysis is to use Bayesian variable selection methods to further investigate the patterns of association across the MHC region.

5.2 Datasets

The data set used for my analysis maps the HLA and non-HLA associations across the entire major histocompatibility complex (MHC) region. An association study on the major histocompatibility complex (MHC) region in SLE using data from the International MHC and Autoimmunity Genetics Network (IMAGEN) study, on 1,199 SNPs from chromosome 6 showed several SNPs with strong evidence of an association. This is a case-control study with 632 UK SLE cases and 746 UK controls from the 1958 Birth cohort [47]; and 483 US SLE cases and 1049 US controls from the New York cancer Project. [48]

In order to increase the power of my analysis to detect separate signals of association with SLE, another dataset was used. The second phase of the IMAGEN study collected data on Spanish subjects. The Spanish dataset has 5,024 SNPs for 813 individuals in the MHC region. This consisted of 404 controls and 409 cases. Combining UK/US and Spanish datasets together results in a larger dataset with more power for statistical analysis but raises concerns about heterogenity. There was an intersection of 777 SNPs between the UK/US and Spanish datasets.

The plot below shows the distribution of SNPs across the MHC region by dataset. The imputed dataset is described in Section5.6.

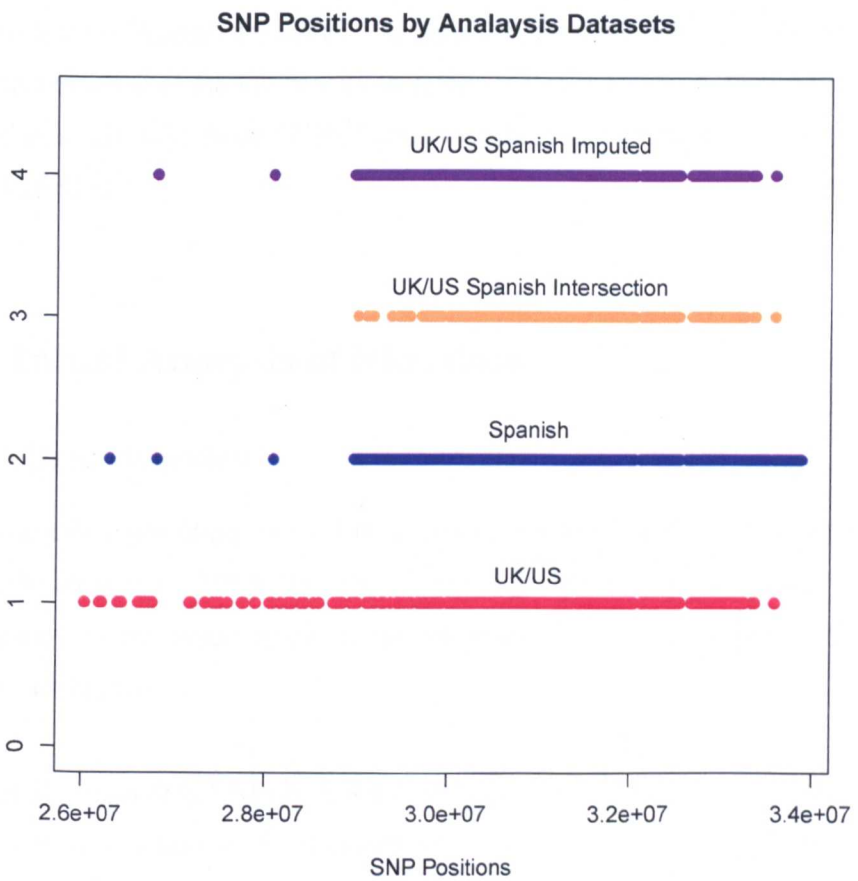


Figure 5.1: Distribution of SNPs by dataset

The following table summarises the final number of SNPs in each dataset after quality controlling (detailed in Section 5.3) and the average marker spacing.

Dataset	Number of SNPs	Cases	Controls	Mean Spacing(BP)	Median Spacing (BP)
UK/US	1,199	1,115	1,795	6,311	1,864
Spanish	5,024	409	404	1,510	409
Imputed	3,592	1,524	2,199	2,135	687

Table 5.1: Average marker spacing by dataset in basepairs

It is clear that the Spanish dataset and the dataset of both UK/US, Spanish and imputed SNPs have much finer marker density than the UK/US dataset (i.e. that used by Rioux et al). This means with more SNPs, I am more likely to find the primary signals associated with SLE.

5.3 Initial Analysis of SLE data

5.3.1 Data Overview

Before quality controlling, the UK/US dataset contained genotype information for 2,921 individuals on 1,230 SNPs. There were 11 people with another family member in the study as determined by the family identification variable. These were dropped from my analysis.

For a details on missing SNPs by UK/US and Spanish data please see Appendix 8.0.1. There were a maximum of 5% of genotypes missing over any one SNP and this minimal missing data was imputed using Mach [17]. The imputation methods used by this program are discussed in 2.7.3. This algorithm uses Estimation Maximisation (EM) to iteratively estimate the missing haplotype probabilities based on the observed haplotypes of each individual at other loci. This method should converge to the haplotype frequencies that equate to the maximum likelihood.

I ran the MACH program for 50 iterations, considering 200 haplotypes at each iteration. This was reasonable for the small amount of missing data [17]. Missing genotype data was imputed into the UK/US dataset using information from the UK/US data, and missing genotype data was imputed into the Spanish dataset using haplotype information from the Spanish data. Missing genotypes were imputed as expected values for each individual.

After imputing the missing genotypes within the UK/US dataset, 14 SNPs were excluded from the analysis because they had the same genotype for every individual. The UK/US data now contained 2910 individuals with 1216 SNPs. This included 632

UK cases, 746 UK controls, 483 US cases and 1049 US controls.

5.3.2 Testing for HWE

I then tested whether or not the SNPs are in Hardy Weinberg Equilibrium (HWE). It is important that SNPs are in HWE because deviations can be a sign of genotyping error, inbreeding, population stratification or selection as mentioned in 2.3.1. Deviations from HWE may invalidate assumptions of the analysis and give incorrect results. Checking for HWE is therefore a necessary data quality check. I used the Pearson goodness of fit test (also known as the χ^2 test) to test for deviations in my control data from HWE. Only the control data is used because if there is an effect of a particular SNP on SLE, for example, then the genotypes for the cases of that SNP will be out of HWE by definition. Note: the alternative HWE test using a likelihood ratio method resulted in the same SNPs being in or out of HWE. Both these tests have 2 degrees of freedom.

This test using UK/US controls showed that 17 SNPs were not in HWE with p-values of less than 10^{-5}. The expected number of SNPs out of HWE in a dataset this size is $<$ 1. These SNPs were not included in my analysis. This left a final dataset for analysis with 1,199 SNPs.

Doing the same test for each of the 5,024 SNPs of the Spanish controls dataset, showed that they were all in HWE with a threshold p-value of less than 10^{-5}.

5.3.3 Population Structure

The population structure between the UK and US datasets was tested to ensure that combining them to form one dataset was sensible. Wrights F_{ST} is a measure of heterozygosity between different populations and tests whether the allele frequencies by SNP in each population are comparable. It was developed by Sewall Wright in the 1920s [49, 50]. The F-statistic can also be thought of as a measure of correlation between genes from different populations. The value of the F-statistic is altered by sev-

eral evolutionary processes, such as mutation, migration, inbreeding, natural selection, but its primary function is to measure the amount of allelic fixation due to genetic drift. An F-statistic of 0 indicates no divergence between populations, and an F-statistic of 1 indicates that the populations are completely different.

The F-statistic between UK and US controls was calculated using the R package polysat [51] to be 0.0006. As described by The International HapMap Consortium [10] and Holsiner & Weir [52] this value is considered to show that the UK and US datasets have similar allele frequencies by SNP and so can be merged to form one dataset with more statistical power for further analyses. Further sensitivity tests to show how analyses changed by UK or US alone will be carried out to illustrate that they obtain similar results. See Section 5.3.4 below and Appendix 8.2.

The F-statistic between the UK/US controls and the Spanish controls was calculated to be 0.005. Although not as close to 0 as the F-statistic between the UK and US controls, this indicates that the allele frequencies by SNP between the two datasets are fairly similar, and merging all the data for one statistical analysis is not unreasonable. [52]

5.3.4 Frequentist test of association

I used a simple frequentist Fisher exact test to initially analyse any SNP associations with SLE as described in 2.8.1 In this framework, each SNP is tested for an association individually. Note: the Bayesian Multivariate Adaptive Regression Spline model I use later 5.9 has a dominance component for flexibility, and assumes an underlying Gaussian distribution of liability (probit model). However, a Fisher exact test is asymptotically equivalent to a logistic model so this makes it slightly more difficult for direct comparisons to be made under the different methods.

UK/US Data Analysis

Applying this test to the 1199 SNPs of the UK/US dataset I found several to have very small p-values. The results are shown in the plot below.

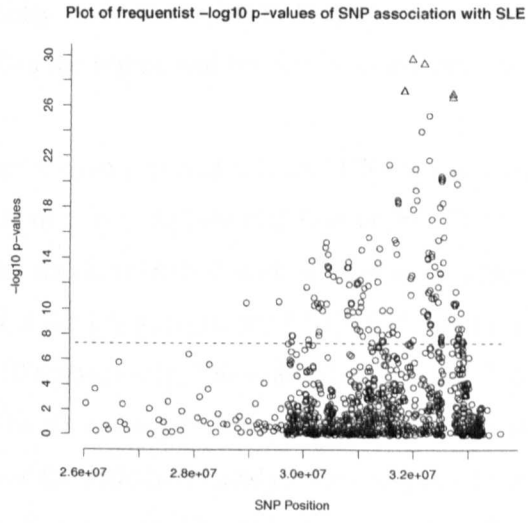


Figure 5.2: Plot of Frequentist p-values of UK/US SNP Association with SLE

This plot shows that (in order of marker position & in left to right on the plot & highlighted by triangles) RS3130484, RS3131379, RS3131378, RS558702, RS1269852, RS2040410, RS2187668 are highly significant, with frequentist p-values of 7.11E-28, 7.11E-28, 7.11E-28, 2.06E-30, 4.54E-30, 2.29E-27, 1.29E-27 respectively. Rioux et al [46] found that RS1269852 had an odds ratio of association with SLE of 2.4 with a p-value of 5.63E-29. The top two SNPs in the above frequentist analysis; RS558702 and RS1269852 (Rioux et al's top SNP) are physically very close to each other (marker positions 31978304 and 32188168) and are in high LD (r^2 of 0.961, D' of 0.985)

Testing a large number of SNPs, it is expected that some SNPs would be significant by chance alone. Therefore in doing genome wide association studies, or in studies with a large number of variables being tested, it is necessary to change the threshold p-value from the "normal" value of 0.05 to one of say $5 * 10^{-8}$. This reduces the false discovery rate of defining too many SNPs to be associated. Using a p-value of $5 * 10^{-8}$ as a cut off (as used by the Welcome Trust Case Control Consortium (WTCCC) [53]),

there are 161 SNPs associated with SLE in the above analysis. The dashed line on the plot represents this cut off.

It is evident from the plot of my simple frequentist analysis that there is a clustered nature of associations across SNPs on chromosome 6. There is no clear signal as to which SNPs are actually associated with SLE. This clustering is likely to be largely due to strong LD within the region and my aim is to disentangle this.

The frequentist p-values from separate UK and US analyses are similar. The top 10 SNPs are the same though in a slightly different order. The top 2 SNPs for the UK analysis are RS558702 and RS1269852 with respective p-values 5.76E-20 and 1.24E-19. The top 2 SNPs for the US analysis are RS1269852 and RS558702 with p-values 3.05E-10 and 5.05E-10 respectively. These are also the same 2 top SNPs from the joint frequentist analysis. In all 3 analyses, the p-values between the top 2 SNPs are very close. The p-values for the individual analyses are slightly bigger but this is likely to be due to the smaller sample sizes. This echoes the result of the F-statistic above, and we can conclude that the datasets are similar enough to combine.

Spanish Data Analysis

Applying the frequentist Fisher exact test as described above to the Spanish dataset, I found several to be highly associated with SLE. The results are shown in the plot below.

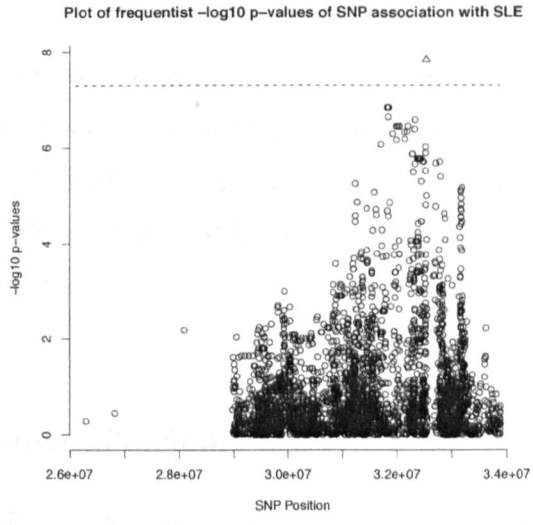


Figure 5.3: Plot of Frequentist p-values of Spanish SNP Association with SLE

Again, there is a clustered nature of associations across SNPs on chromosome 6. There is no clear signal as to which SNPs are independent signals associated with SLE. Note that there is less power in this analysis than in the UK/US one as there is a smaller sample size (813 individuals vs 2920 individuals).

Using p-values of $5 * 10^{-8}$ as a threshold, there is only one SNP associated with SLE in the Spanish dataset. This SNP is marked as a triangle on the plot above and is RS9268832 with a p-value of 1.46E-08. However, the plot shows that (in order of marker position and from left to right on the plot), RS3131381, RS3131379, RS3117574 an RS3130490 all share a p-value of 1.45E-07 and are located very close to each other and the top ranking SNP on chromosome 6 (marker position numbers 31816442, 31829012, 31833209, 31847099. The top SNP has marker position 32535767) so it is likely that these are all highly correlated with each other. The next stage was therefore to investigate the LD between all these SNPs to see if this is the case.

Note: the top two SNPs from the Spanish analysis are not in the UK/US dataset but the third most highly associated SNP RS3131379 is the fourth highest in the UK/US dataset. SNP RS3131379 has a p-value of association with SLE in the UK/US dataset of 7.11E-28 and 1.45E-07 in the Spanish analysis. The difference in p-values could be due to significantly less power in the Spanish dataset due to a smaller sample size.

The top SNP RS558702 from the frequentist UK/US analysis with a p-value of 2.06E-30 has a p-value of 3.52E-07 in the Spanish analysis. The second highest SNP from the UK/US dataset, RS1269852, has a p-value of 4.54E-30 in the UK/US analysis and 4.50E-07 in the Spanish one. These SNPs are ranked 8th and 14th respectively in the Spanish frequentist analysis.

5.3.5 Linkage Disequilibrium

UK/US Data Analysis

I then investigated the LD between the top ranking 100 SNPs from the UK/US frequentist analysis. An LD plot created using Haploview [54] [55] shows the levels of LD between all SNPs selected in order of marker position on chromosome 6 from left to right. The darker the box connecting the two SNPs, the higher the level of LD between the two SNPs. The LD value shown in each box is the r^2 statistic. The black boxes drawn on the plot show blocks of LD.

The LD plot of all UK/US SNPs (in Appendix 8.1) shows that all of these SNPs are in high LD with several other SNPs. This confirms that the clustering shown in the frequentist results above is likely to be due to strong LD amongst the SNPs.

The levels of LD between the 10 highest ranking SNPs from the UK/US frequentist analysis are shown in the plot below. They are all in LD with at least one other SNP in the top 10 with $r^2 > 0.5$.

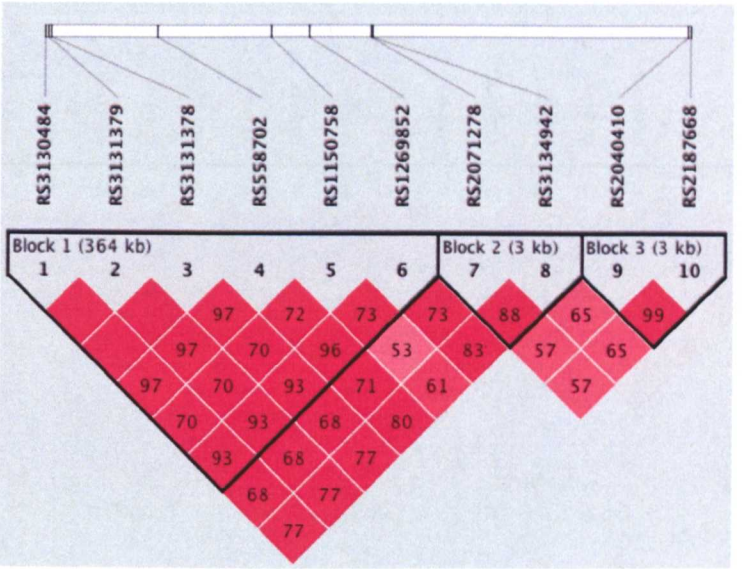


Figure 5.4: LD Plot of Top 10 UK/US SNPs from Frequentist Test of Association with SLE

This shows that the top two SNPs in the above analysis (RS558702 and RS1269852) are highly correlated with all top SNPs showing an association. The very high correlation between these SNPs could mean that there is only one locus associated with SLE but it is difficult to establish which one.

Spanish Data Analysis

I then investigated the LD between the top ranking 100 SNPs from the Spanish frequentist analysis. The LD plot (in Appendix 8.1) shows the levels of LD between all the SNPs.

To see more clearly, the levels of LD between the 20 highest ranking SNPs from the frequentist analysis are shown in the plot below. The top 20 were chosen rather than the top 10 in the UK/US dataset because the loci in the Spanish dataset are more dense. I also wanted to include the two top SNPs from the UK/US for comparison. Again, the darker the shade of red the box connecting the two SNPs, the higher the level of LD. The top 20 SNPs are all in LD with at least one other SNP with $r^2 > 0.7$ as shown below.

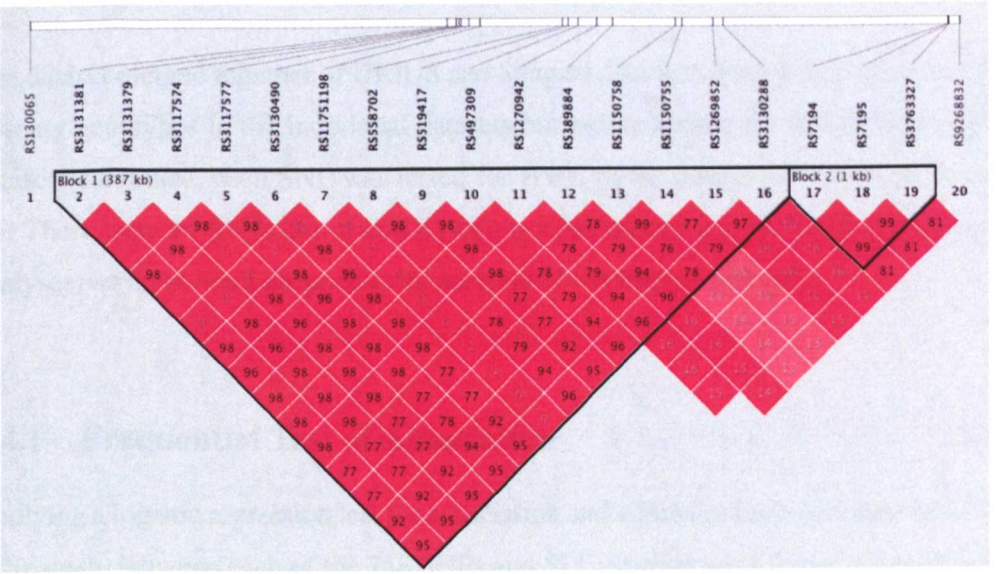


Figure 5.5: LD Plot of Top 20 Spanish SNPs from Frequentist Test of Association with SLE

5.4 Intersection of UK/US and Spanish Data

In order to increase power and the ability to localise the signals of association, the UK/US and Spanish datasets were merged. Combining the UK/US datasets and the Spanish datasets resulted in 772 overlapping SNPs for 3723 individuals; 1,524 cases and 2,199 controls.

It was necessary to flip alleles on 230 SNPs so that those from the Spanish dataset were on the same strand as those from the UK/US dataset. Sometimes if the genotyping of the SNPs are done on different chips e.g. Affymetrix and Illumina, then the major and minor alleles can be defined differently. This means that Cs are coded as Gs, or Ts are coded as As or vice versa. This was the case between the Spanish and UK/US datasets for 230 SNPs, and so I flipped these strands on the Spanish dataset to match those on the UK/US dataset.

There were 34 SNPs with different observed minor allele frequencies. These were genotypes coded across both studies as C/G or A/T but had different minor alleles. In order to reconcile this problem, I compared the allele frequencies to those in the European HapMap database and matched the minor allele to this information. [56]

The dataset merged together of UK/US and Spanish data was done after imputation for missing genotypes in the individual datasets but before testing for HWE. Using both datasets combined, each SNP was tested for HWE in the dataset with more individuals. There were 14 SNPs out of HWE (with p-value criteria as in UK/US and Spanish analyses) so these were dropped from analysis of the intersection data.

5.4.1 Frequentist Test of Association

Applying a logistic regression test for association and adjusting for population (UK/US vs Spanish) between each of the 758 SNPs and SLE, several were found to be strongly associated. The results are shown in the plot below.

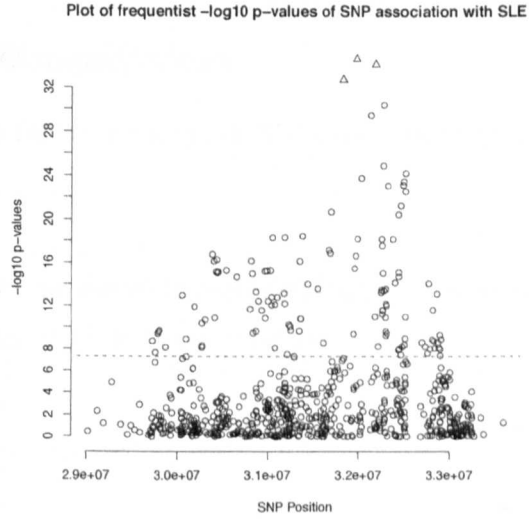


Figure 5.6: Plot of Frequentist p-values of Intersection SNP Association with SLE

The plot above shows that the three SNPs with the highest frequentist p-values after adjusting for population (UK/US vs Spanish) for an association with SLE (in order of marker position and from left to right, and marked by triangles) are RS3131379, RS558702 and RS1269852 with respective p-values of 1.97E-33, 2.73E-35 and 8.34E-35. The top SNP RS558702 is the top SNP in the UK/US individual frequentist analysis (p-value of 2.06E-30), and is the 8th highest in the Spanish analysis (p-value of 3.52E-07). The second most significant SNP RS1269852 ranks 2nd in the UK/US analysis (p-value 4.54E-30) and 14th in the Spanish analysis (p-value 4.50E-07). Using a p-value threshold of a conservative $5 * 10^{-8}$, there are 124 SNPs in the overlapping dataset that have an association with SLE.

Again, there is a clustered nature of associations across SNPs and it is not obvious which SNPs are actually associated with SLE.

5.4.2 Linkage Disequilibrium

The LD plot between the top ranking 100 SNPs from the frequentist analysis is shown in Appendix 8.1.

The levels of LD between the 10 highest ranking SNPs from the frequentist analysis are shown in the tables of D' and then r^2 below.

Table 5.2: Measures of LD, D', Between Top 10 SNPs from Frequentist Analysis

	RS558702	RS1269852	RS3131379	RS3134942	RS1150758	RS2071278	RS2395171	RS2227139	RS3135366	RS2239805
RS558702										
RS1269852	0.985									
RS3131379	0.990	0.975								
RS3134942	0.955	0.977	0.937							
RS1150758	0.987	0.994	0.967	0.839						
RS2071278	0.957	0.980	0.938	1.000	0.725					
RS2395171	0.911	0.932	0.897	0.790	0.686	0.758				
RS2227139	0.929	0.951	0.914	0.774	0.722	0.724	0.995			
RS3135366	0.913	0.933	0.900	0.801	0.689	0.760	0.982	0.973		
RS2239805	0.910	0.931	0.897	0.800	0.686	0.756	0.976	1.000	0.974	

Table 5.3: Measures of LD, r^2, Between Top 10 SNPs from Frequentist Analysis

	RS558702	RS1269852	RS3131379	RS3134942	RS1150758	RS2071278	RS2395171	RS2227139	RS3135366	RS2239805
RS558702										
RS1269852	0.959									
RS3131379	0.972	0.931								
RS3134942	0.791	0.818	0.767							
RS1150758	0.737	0.739	0.714	0.615						
RS2071278	0.681	0.705	0.660	0.858	0.516					
RS2395171	0.568	0.587	0.556	0.505	0.425	0.529				
RS2227139	0.197	0.204	0.192	0.158	0.157	0.161	0.330			
RS3135366	0.570	0.587	0.557	0.505	0.428	0.531	0.961	0.316		
RS2239805	0.554	0.573	0.542	0.494	0.416	0.515	0.931	0.341	0.930	

Considering the plot of frequentist p-values of the overlapping SNP analysis and the LD between those SNPs with the top10 highest p-values, we may believe that there could be two underlying causal loci. One represented by the two top hits RS1269852 and RS558702 which are in extremely high LD ($D' = 0.99$), and the other represented by the 4 in the cluster above in the frequentist plot of p-values (RS2239805, RS3135366, RS2395171 and RS2227139). These 4 SNPs are in very tight LD with each other ($D' > 0.97$). This would collaborate with evidence suggested by the single analyses but gives a slightly more detailed picture.

5.5 Summary of Frequentist Analysis

The table below gives a comparison of the top 10 SNPs from intersection frequentist analysis with results for the same SNPs in individual UK/US and Spanish analyses.

Table 5.4: Summary of p-values of Association & HWE p-values by Top SNP & Dataset

	Intersection Data	UK/US Data	Spanish Data
SNPs	Association p-val (HWE p-val)	Association p-val (HWE p-val)	Association p-val (HWE p-val)
rs558702	3.13E-31 (0.18)	2.06E-30 (0.24)	3.52E-07 (0.35)
rs1269852	8.68E-31 (0.09)	4.54E-30 (0.12)	4.50E-07 (0.36)
rs3131379	3.24E-29 (0.40)	7.11E-28 (0.51)	1.45E-07 (0.35)
rs3134942	4.59E-27 (0.76)	6.37E-26 (0.82)	1.35E-06 (0.16)
rs1150758	5.32E-25 (0.06)	1.09E-24 (0.15)	4.65E-07 (0.70)
rs2071278	2.65E-22 (0.76)	2.16E-22 (0.88)	3.41E-04 (0.08)
rs2395171	1.1E-20 (0.63)	8.72E-21 (0.69)	1.11E-03 (0.38)
rs2227139	1.64E-20 (0.56)	4.96E-19 (0.63)	1.29E-06 (0.36)
rs3135366	1.64E-20 (0.84)	4.05E-21 (0.91)	3.44E-03 (0.51)
rs2239805	4.58E-20 (0.21)	6.16E-21 (0.52)	7.89E-03 (0.42)

From this it is clear that the p-values **within** each dataset analysis are similar. This

makes it difficult to distinguish the SNPs that are independent signals.

I then carried out simple frequentist tests of association by conditioning on the top SNPs RS558702 and then RS1269852 in order to examine if there are any independent signals within the top 10 SNPs from the intersection analysis.

Table 5.5: Table of p-values of Association Conditioning on RS558702

SNPs	p-value
rs1269852	0.46
rs3131379	0.18
rs3134942	0.23
rs1150758	0.50
rs2071278	0.66
rs2395171	0.31
rs2227139	6.84E-06
rs3135366	0.33
rs2239805	0.37

Table 5.6: Table of p-values of Association Conditioning on RS1269852

SNPs	p-value
rs3131379	0.62
rs3134942	0.30
rs1150758	0.41
rs2071278	0.80
rs2395171	0.37
rs2227139	7.78E-06
rs3135366	0.37
rs2239805	0.43

These results show that there are at least two independent signals. It would be more

efficient to test for an association with all SNPs jointly. This is done in Section 5.9

5.6 Combined Dataset with Untyped SNPs from HapMap

In order to obtain a better coverage of the whole chromosome and therefore to be more likely to find independent SNPs, untyped genotypes in each dataset were imputed using information from HapMap (as discussed in Section 2.7.3).

The SNPs missing from each dataset but available in the other, and in HapMap were imputed in two blocks. Firstly, those missing from the UK/US but in the Spanish dataset were imputed using information from the UK/US dataset and HapMap. Then those missing from the Spanish dataset were imputed in the same way. This took into account the fact that the two datasets might have different population structures, and imputed SNPs into the the Spanish dataset, for example, using information from the UK/US dataset would not make sense if they were different. In fact, Wright's F_{ST} between the UK/US and Spanish datasets for those SNPs in both datasets is 0.005. This implies that they datasets have similar allele frequencies. The imputed datasets were combined and an indicator variable was added to determine between UK/US and Spanish to allow for different MAFs or effect sizes in the two datasets.

The SNPs used to impute untyped SNPs in each of the datasets were taken from HapMap. These SNPs were unrelated Utah residents with ancestry from northern and western Europe. The SNP samples were taken from chromosome 6 within the MHC region (26000000 to 34000000) for 17 individuals each. Mach [17] was again used for imputation using 50 iterations using information from 200 haplotypes for each SNP at each iteration. This resulted in a dataset of 3,636 SNPs for 1,524 cases and 2,199 controls. 2,910 individuals from UK/US (1,795 controls and 1,115 cases) and 813 individuals from the Spanish data (404 controls and 409 cases). I used the expected genotype value output from Mach so SNPs are now 1.2, 0.7, etc., for example. I filtered any imputed SNPs using r^2, which estimates the squared correlation between imputed and true genotypes. I used a cut-off of 0.8. [56]. This resulted in a dataset for analysis with 2,733 SNPs.

The tight LD of the associated genotypes motivates the use of a model search method, for which Bayesian methods cope better with uncertainty about the model.

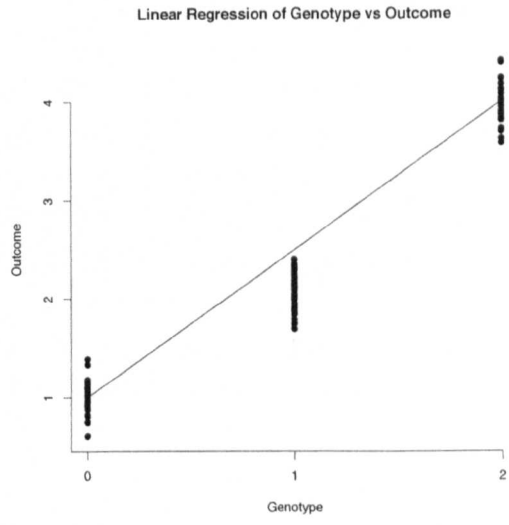
5.7 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Spline (MARS) modelling was developed by Friedman [57] to allow for flexible regression of high dimensional data. This model was motivated by the fact that it can be difficult to approximate the relationship between an outcome and many variables and we may not know a priori what effect we expect each variable to have upon the outcome. The set-up of MARS models is described in detail below.

In a genetic context, a MARS model does not force a specific model for each locus. It allows different SNPs to have different effects on the outcome of interest. For example, some SNPs may be dominant while others may be additive or recessive. MARS models also account for non-linear relationships between outcome and variable, can allow for interactions, and use variable selection to include the most significant variables in the model.

5.8 Non-linear regression

Sometimes we want to allow for non-linear relationship of the genotypes of a SNP with an outcome. For example, if we were to fit a linear model to this data.

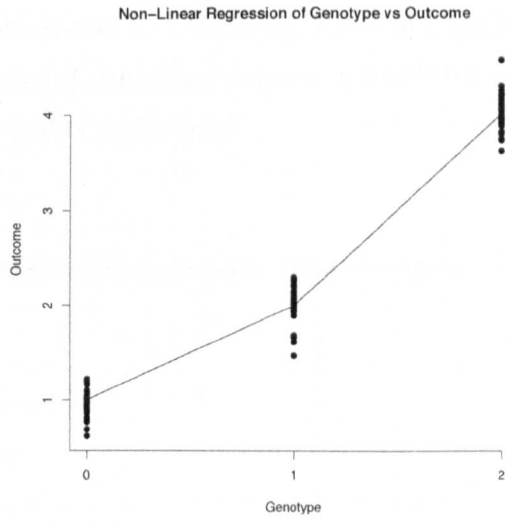


$$\hat{y} = 1 + 1.5x \tag{5.1}$$

From the plot, it is clear that this is not a linear relationship. A better fitting model would be

$$\hat{y} = 1 + 1[x - 0]_+ + 1[x - 1]_+ \tag{5.2}$$

as shown in the plot below. Where $[x - 1]_+$ is known as a basis function, and $[]_+$ is the value of that in the brackets if it is positive; 0 otherwise. For example, $[2 - 1]_+ = 1$ but $[2 - 3]_+ = 0$. In the example above, the gradient changes at 1. This is known as a knot.



This is an example of a simple MARS model.

For a single covariate, x, a regression spline model with basis functions for varying knot points, can be written as

$$g(x, \beta, t, K) = \beta_0 + \sum_{i=1}^{K} \beta_i [x - t_i]_+^q \tag{5.3}$$

where q is a positive integer to denote the order of the spline (e.g. q=2 for a quadratic spline model; q=1 in these examples), as mentioned before $[]_+$ is the positive part of that in the brackets, $t_1, \ldots, t_K$ is a set of candidate knots. Note that for genetic data, coded as 0, 1 or 2, we only allow knots at values 0,1 or 2. To reiterate, knots are points on the x-axis where the nature of the function changes. All these parameters are estimated simultaneously with the regression coefficients β and K, the total number of knots.

As mentioned above, three well known genetic models are the dominant, recessive or additive model on an outcome. When modelling a disease outcome and we want to model the probability of disease/ no disease (1 / 0) as for SLE outcome, it is desirable

to work using the probit model on a log scale. This was explained in detail in Section 2.8.2 However, examples of dominant, recessive and additive models are plotted without a log scale for simplicity at this stage.

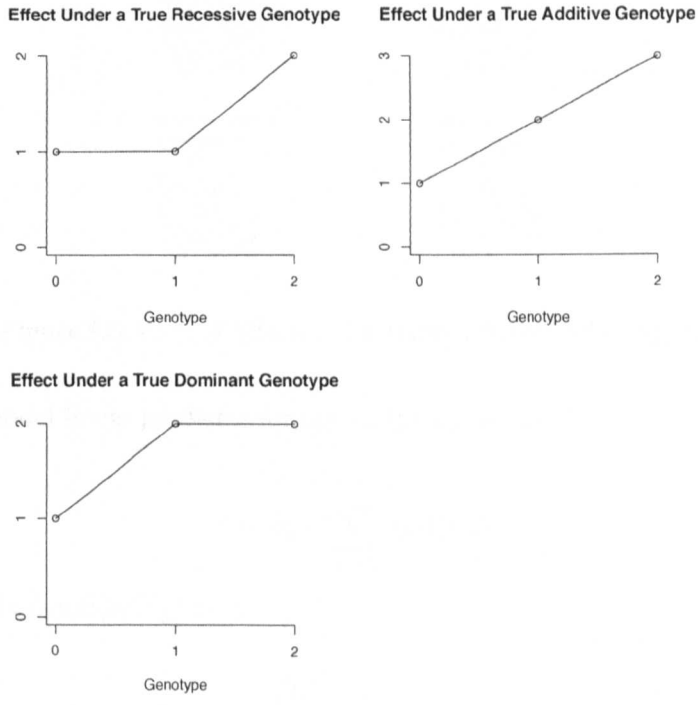


Figure 5.7: Plots of Effects of True Recessive, Dominant and Additive Genotypes

From these plots, it is clear that MARS model is suitable to fit to genotype data, for flexible models, including those usually fitted, and can allow a different model for each locus.

The MARS model can be extended to include interaction terms. This can be useful if for example, genotypes have an interaction effect. For example, if we have an interaction between true additive, dominant and recessive SNPs, the effects will look like the plots below.

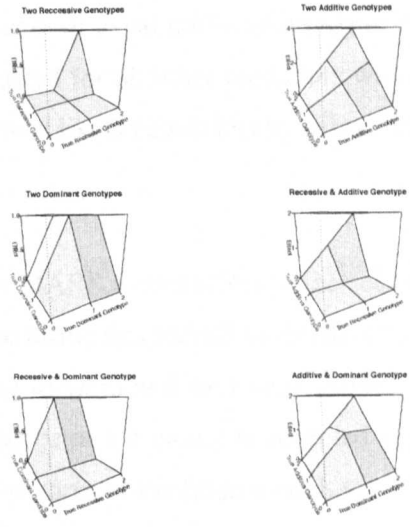


Figure 5.8: Plots of Effects of Interactions Between Genotypes

The MARS model linear predictor for my analysis is written

$$\eta = \beta_1 + \sum_{k=2}^{K} \beta_k \mathbf{B_k}(\mathbf{x}) \tag{5.4}$$

where the basis function $\mathbf{B_k}(\mathbf{x})$ is

$$\mathbf{B_k}(\mathbf{x}) = \prod_{\mathbf{z=1}}^{\mathbf{z_j}} [\mathbf{s_{kz}}(\mathbf{x_{v(kz)}} - \mathbf{t_{kz}})]_+ \tag{5.5}$$

where z_j=1or 2,... depending on number of interactions for that particular basis function, $[.]_+$ = max(0,.), z is the degree of interaction of basis function $\boldsymbol{B}_k$, $s_{kz} = -1$ or 1 depending on the sign (Note: changing the sign of the basis function can lead to effects by genotype opposite to those shown on the plots above], v(kz) is a member of 1,...,p and indicates the predictor in the model and t_{kz} is the associated knot value with interaction term z for basis function k. Therefore, η is a linear combination of non-linear transformations of the covariates. Let $\hat{\boldsymbol{\theta}} = \left\{\hat{\mathbf{k}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{z}}, \hat{\mathbf{s}}, \hat{\mathbf{v}}, \hat{\mathbf{t}}\right\}$ be a vector of parameter estimates. In this equation each predictor is constrained to appear only once in each basis function.

In this analysis for SLE, the MARS model is restricted to have a maximum of 2 interactions.

Notice that, just as in the case of usual multivariate regression, the effect of each predictor in the model is adjusted for all other predictors within the model. This should enable us to tease out the most likely causal loci by automatically correcting for nearby associated SNPs.

The 'Adaptive' part of the MARS model refers to the selection of the optimal model. The frequentist approach to fitting this MARS model iterates through the possible models (basis functions), using forward and backward variable selection. As in stepwise regression, a proposed change in the model is accepted if it results in a significantly improved residual sum of squares of the fitted model.

In forward selection, the candidate spline term multiplied by the existing basis function that gives the largest reduction in residual sum of squares is added. In order to reduce the number of basis functions in the model and to avoid overfitting, a backward deletion is proposed. We can choose which basis functions to delete using the generalised cross-validation criterion, for example. [58]

5.9 Bayesian Multivariate Adaptive Regression Spline (BMARS)

It is desirable to search over all "possible" models in order to find those SNPs that are most important in the prediction of SLE. The BMARS model used should automatically correct for nearby associated SNPs, and only those most directly associated should be included in the model.

I used a Bayesian Multivariate Adaptive Regression Spline (BMARS) model, developed by Verzilli et al. [59], to identify the most associated SNPs taking into account nearby associated SNPs in the data above via Bayesian model averaging.

A Bayesian approach summarises the evidence in favour of model m in terms of the posterior probability of model m given the data ($f(m|y)$). [60]

$$f(m|y) = \frac{f(y|m)f(m)}{\sum_{m \in M} f(y|m)f(m)} \tag{5.6}$$

Model averaging refers to averaging over all possible models using either

$$a) f(\beta|y) = \sum_{\text{all models m including}\beta} f(\beta|y,m)f(m|y) \tag{5.7}$$

or

$$b) p(\text{covariate i being in model}) = \sum_{\text{all models including covariate i}} p(m|y) \tag{5.8}$$

Verzilli et al. use a reversible jump algorithm as described in Section 3.7 [34,38]. This allows the MCMC scheme to sample from any model m for the MARS models considered. The reversible jump algorithm explores the space of Θ, proposing to change the dimension of θ at each iteration using a birth, death or switch step.

For the MARS models used in this analysis, the acceptance probability of a new basis function (birth step) is simply [61]

$$min\left\{1, BF(\theta', \theta)R\right\} \tag{5.9}$$

where R is the ratio of probabilities ($\frac{d_{k+1}}{b_k}$ as described in Section 3.7) and BF((θ', θ) is the Bayes factor (see Section 3.6) of the proposed model (θ') compared to the current model (θ).

5.10 Priors for Parameters in the SLE BMARS Model

I follow Verzilli et al. who use similar prior distributions for the parameters $\theta = \{\mathbf{k}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{s}, \mathbf{v}, \mathbf{t}\}$ as Holmes and Denison 2003. Using a Bayesian approach, $\hat{\theta}$ is treated as unknown, and all parameters are assigned prior distributions. The prior distribution on the sign indicator for basis function k, s_{kz} is uniform on $\{-1, 1\}$, i.e. p(s_{kz})=U(-1,1). Predictor variable v(kz) which is used in term z of basis function k and indicates whether a SNP is included in the model or not, is given a uniform prior p(v(kz))=U(0,...,p) where p is the number of possible SNPs in the model. The knot values t_z are uniformly distributed on the observed genotype values, i.e p(t_{kz})=U(0,1,2). For the maximum number of basis functions K that the model is allowed to grow to we set $p(K) = U(1, \ldots, K_{max})$, here choosing K_{max}=250. The prior distribution for the vector of spline coefficients, β, is $p(\beta) = MVN(\mathbf{0}, \sigma_\beta^2 \mathbf{I})$. Finally, the prior for σ_β^2 is inverse gamma, i.e. $p(\sigma_\beta^{-2}) = gamma(0.01, 1)$.

The BMARS code was then extended by Verzilli et al. to include a Poisson prior on the number of variables in the model, as previous code assumed at least one SNP associated with SLE at each iteration. $p(K) \sim Pois(\lambda)$

λ was set to 0.5 as this equates to a very conservative prior of less than one SNP being included in the model. However, under sensitivity analysis, by varying the values of λ, neither the posterior probability or the number of SNPs being included changed.

In order to simplify sampling from posterior distributions, the probit link function with data augmentation was used, as explained in Section 2.8.2. The advantage of using the latent variables w_i, together with conjugate priors, is that posterior sampling of all parameters is simplified following from the Bayesian linear model, conditioned upon w_i.

5.11 Application to SLE association study data

In the SLE dataset, observed values of Y are defined by

$$y = \begin{cases} 0 & \text{if no SLE i.e. control} \\ 1 & \text{if have SLE i.e. case} \end{cases}$$

The matrix of $\mathbf{x}$ is made up of SNPs, coded as 0,1,2 for the different possible genotypes as described above in the frequentist analysis; or by expected genotype values in the case of imputed SNPs.

In every case the prior probability is set so that the expected number of SNPs in the model is 0.5. Sensitivity of this was examined by altering the mean of the prior ($p(K) \sim Pois(\lambda)$) from 0.5 to 10 as this could be a potentially informative prior (see plot below).

The plot below of probability mass functions of the Poisson distribution with λ = 0.5, 5,10 shows that this could be a potentially informative prior on the number of SNPs in the model.

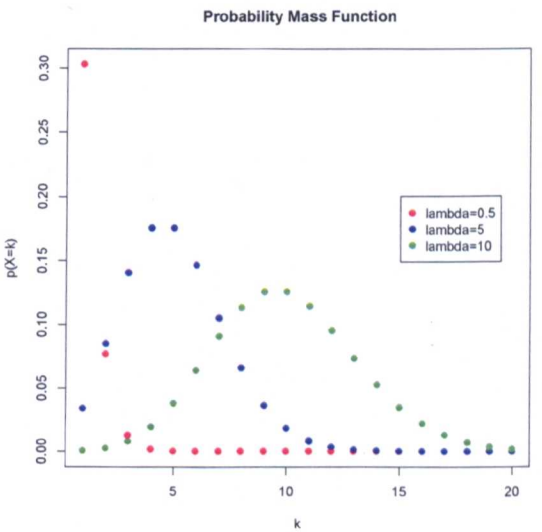


Figure 5.9: Probability Mass Functions of the Poisson Distribution

5.11.1 Analysis of UK/US dataset

The BMARS MCMC algorithm developed by Verzilli et al. was applied to the UK/US dataset with 1,199 SNPs for 2,910 individuals. The algorithm was run 10 times for 5,000,000 iterations with a burn-in period of 200,000 with a thin of 800. A large thinning parameter was used to reduce the size of the vector stored in R over 5 million iterations. A convergence plot of the posterior probability of the SNP with the highest signal shows that only the short burn-in used is required. The model converges at 150,000 iterations but a burn-in period of 200,000 was used to be stringent.

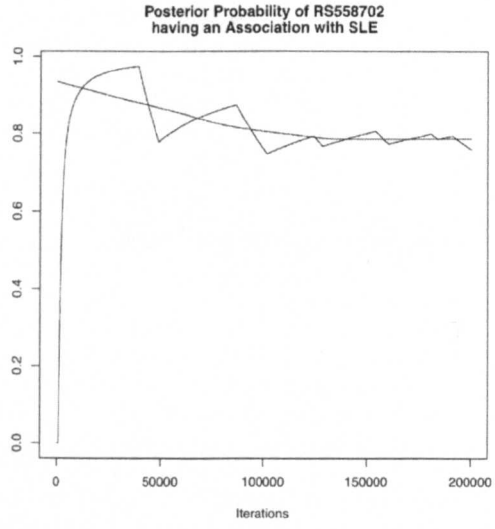


Figure 5.10: Cumulative Posterior Probability of RS558702 having an Association with SLE

The BMARS algorithm was run on the UK and US datasets separately to examine whether the results would be that different to each other. The posterior probabilities of each SNP having an association with SLE were similar. See Appendix 8.2

The BMARS algorithm was run on the combined UK/US dataset for 5 million iterations with a burn in period of 200,000, and thinning of 800, and repeated 10 times. The posterior probability of each SNP having an association was estimated by the number of times the SNP was in the model over all 10 runs divided by 60,000 (the number of iterations stored). The algorithm was run 10 times in order to check whether there was any heterogenity between runs. In each case, the posterior probability of each SNP being included in the model was approximately equal. The posterior probabilities of each SNP are shown in the plot below.

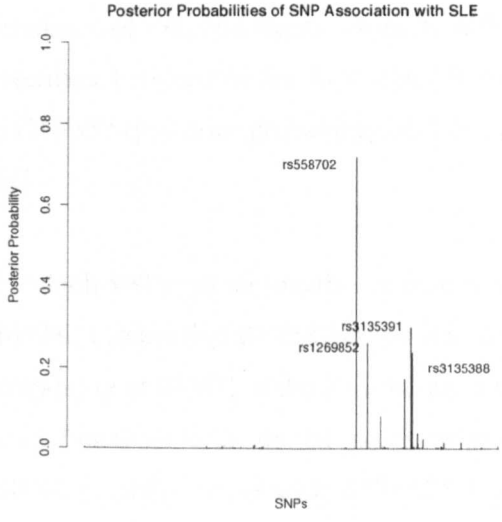


Figure 5.11: Posterior Probabilities by SNP of Association with SLE

This plot shows that SNP RS558702 has the highest posterior probability of 0.72 of having an association with SLE. This was also the top SNP in the UK/US frequentist

analysis with a p-value of 2.06E-30. The 2nd ranked SNP in the frequentist analysis, RS1269852 (p-value 4.54E-30), has a posterior probability of 0.26. This leads to the conclusion that RS558702 has the primary association with SLE.

Note: The posterior probability of 0.72 of RS558702 having an association with SLE does not seem that high compared to a frequentist p-value of association of 2.06E-30. The weaker Bayesian result could be due to the possibility that there is more than one SNP effect on SLE and so with a **joint** analysis, individual effects are smaller as it is less clear which SNPs have an association. The frequentist p-values are from independent tests of association with SLE and do not take into account the joint effects or LD.

Analysing the number of basis functions in the model, when RS558702 is in the model, we found that there were was a probability of 0.12 of only RS558702 being in the model, a probability of 0.74 of there being 2 in the model, and probabilities of 0.13 and 0.01 of 3 and 4 basis functions being in the model with RS558702 respectively. From this we can conclude evidence of a second signal. From the above posterior probability plot, it is difficult to determine between the levels of signal from RS3135391 (posterior probability of 0.3) RS1269852 (posterior probability of 0.26) and RS3135388 (posterior probability of 0.24).

Examining further into which SNPs appear together in each model when any of the top four SNPs are in the model, I found that RS558702 and RS1269852 are very rarely in the model together (probability of 0.0001 of the other being in the model when one is). When there are two basis functions in the model (posterior probability of 0.73 of that being the case), RS558702 is in the model with RS3135391 with probability 0.24 or RS3135388 with probability 0.12. RS1269852 is in the model with RS3135391 with probability 0.09 or RS3135388 with probability 0.07.

The prior probability of having 2 basis functions in the model is 0.08.

There do not appear to be interactions of SNPs in the models with the top posterior probabilities. The probability of RS558702 having an interaction with any other SNP is 0.05.

Frequencies (out of the number of iterations with a model using each SNP) for knot values (i.e. where the gradient of the basis function changes) were plotted for each of the top SNPs associated with SLE from the BMARS output to show the type of relationship (dominant, recessive, additive, for example).

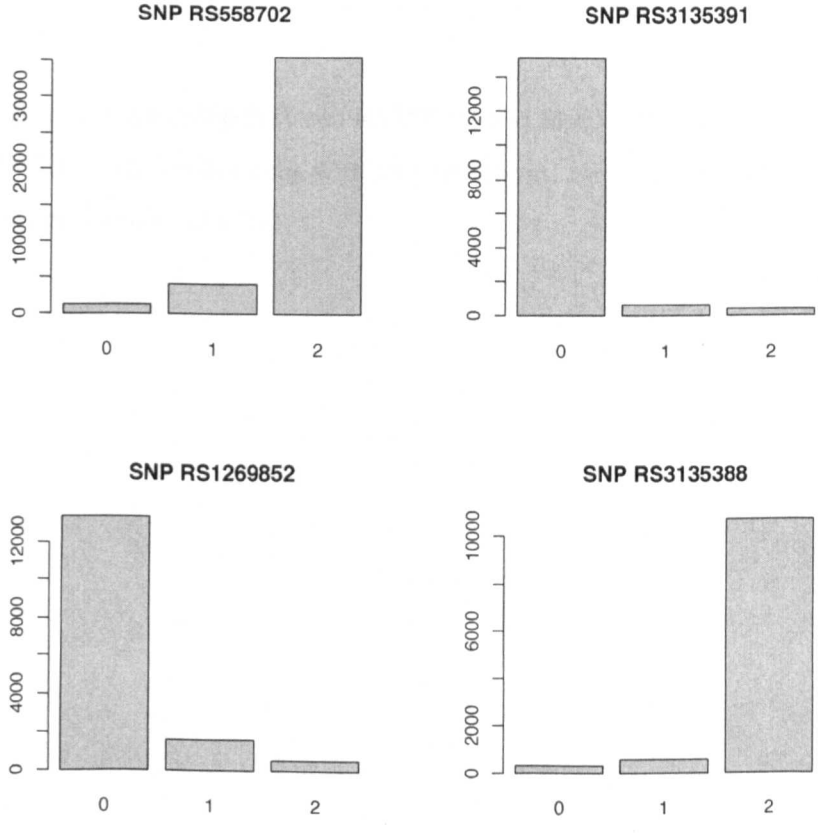


Figure 5.12: Frequencies of Knot Points by Top SNPs

This plot shows that for the top SNP, RS558702, the relationship with SLE appears to be additive. The sign of the equation is altered by s, so we are interested in the values of s against the knots, t (as well as the sign of β). s is always 1 (positive) when the knot value is 0, and always -1 (negative) when the knot value is 2. It can be either positive or negative when the knot is at 1. Given the knot value for RS558702 is 2, this results in the following basis function values at each genotype value

Table 5.7: Basis Function by Genotype (x) for RS558702

x	Basis function:$[-(x-2)]_+$
0	2
1	1
2	0

This leads to a relationship between RS558702 and SLE on the probit scale of (this is for RS558702 in the model only, with no interactions, and a knot point of 2 which has a posterior probability of 0.07)

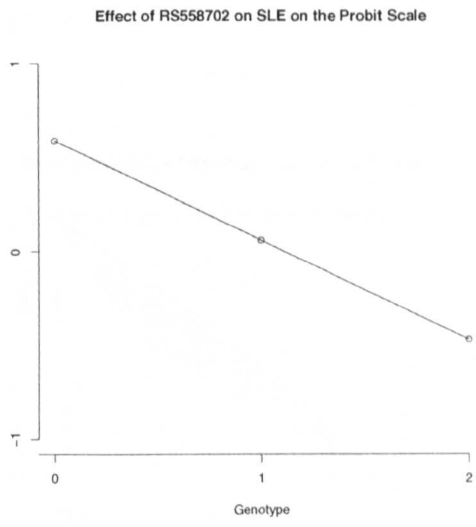


Figure 5.13: Plot of Relationship of RS558702 on SLE

$$\eta = -0.47 + 0.53([-(x-2)]_+) \tag{5.10}$$

where the parameter estimates are the posterior means given the model.

The model with the highest posterior probability (0.13) involves two basis function with SNPs RS558702 (x_1) and RS3135391 (x_2) with no interactions. The basis function involving RS558702 has an s of -1 and a knot point (t) of 2. The basis function for RS3135391 has an s of 1 and a knot point of 0. The effect on the probit scale of this model is

$$\eta = -0.552 + 0.56([-(x_1-2)]_+) + 0.24([(x_2-0)]_+) \tag{5.11}$$

A plot of this model looks like

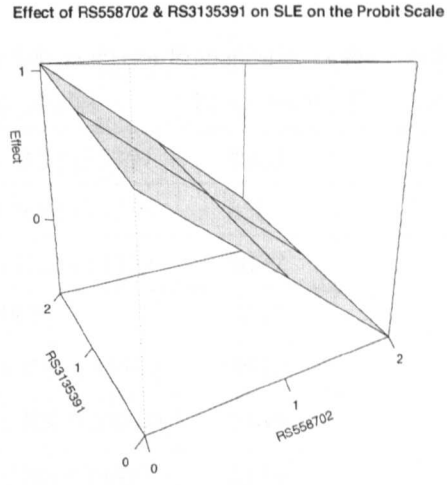


Figure 5.14: Plot of Effect of Most Common Posterior Model on SLE

Below is a table of posterior probabilities of the most common combination of SNPs. Note: These model frequencies are irrespective of knots.

Table 5.8: Posterior Probabilities of the Top Models

SNPs in Model	Frequency	Posterior Probability
RS558702 + RS3135391	8921	0.15
RS558702 + RS3135388	6568	0.11
RS558702 + RS3135352	5032	0.08
RS558702	5017	0.08
RS1269852 + RS3135391	3532	0.06
RS1269852 + RS3135389	2345	0.04
RS558702 + RS396960	2214	0.04
RS1269852	2126	0.04
RS1269852 + RS3135352	1769	0.03

Note: RS396960 (position 32299558) and RS3135352 (position 32500883) have marginal posterior probabilities of 0.08 and 0.17 respectively.

Below are tables of LD measure D' and then r^2 of the top SNPs and those included in the most frequent models.

Table 5.9: D' Between Top SNPs

	RS558702	RS3135391	RS3135388	RS1269852	RS396960	RS3135352
RS558702						
RS3135391	0.823					
RS3135388	0.820	1.000				
RS1269852	0.985	0.822	0.820			
RS396960	0.953	0.930	0.930	0.953		
RS3135352	0.819	0.997	0.997	0.819	0.930	

Table 5.10: r^2 Between Top SNPs

	RS558702	RS3135391	RS3135388	RS1269852	RS396960	RS3135352
RS558702						
RS3135391	0.022					
RS3135388	0.022	0.996				
RS1269852	0.961	0.022	0.022			
RS396960	0.053	0.049	0.049	0.052		
RS3135352	0.022	0.989	0.993	0.022	0.049	

Although, it appears in the plots above that there are 2 clear signals of an association with SLE (Plot 5.11 and Plot5.8); most likely at RS558702 and RS3135391, the SNPs in the top models are in high LD. RS558702 and RS1269852 are in high LD with

r^2=0.96; and RS3135391 and RS3135388 are in high LD with r^2=1.

Note: The r^2 values indicate less LD between the top SNPs than D'. As described in Section 2.5 this is a sign that the MAF varies between the SNPs.

In addition, the mean values of β for RS558702 and RS1269852 are similar (0.545 and 0.538). The same is true for the mean values of β for RS3135391 and RS3135388 (0.245 and 0.247). For plots of posterior densities of β coefficients of top SNPs given they are in the model please see Appendix 8.3. This implies that RS558702 and RS1269852, and RS3135391 and RS3135388 have similar effects on SLE when in the model.

However, as the model has not selected between these 2 pairs of SNPs in high LD, it could be possible that these top SNPs are due to a single underlying untagged locus. Therefore, I aim to impute untyped SNPs and combine this data with another dataset to provide more information.

5.11.2 Analysis of Spanish dataset

The BMARS MCMC algorithm described above was applied to the Spanish dataset with 5,231 SNPs for 813 individuals. The code was run 10 times for 5 million iterations with a burn in period of 500,000, and thinning of 800. The BMARS model of the Spanish dataset took longer to converge than the analysis of the UK/US dataset. This is likely to be due to the larger number of SNPs in the Spanish dataset for model selection (5,231 in the Spanish dataset compared to 1,199 in the UK/US). A convergence plot of the posterior probability of the SNP with the highest association with SLE is shown below.

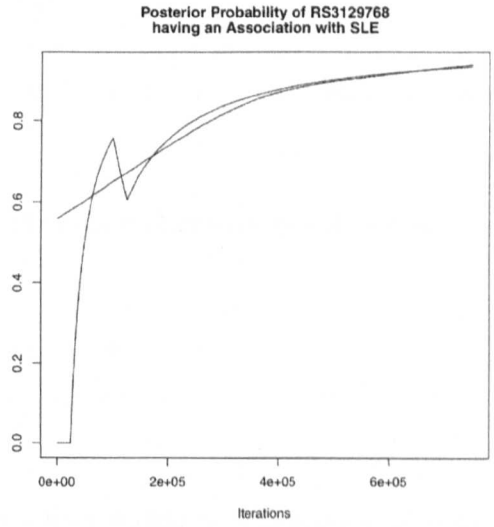


Figure 5.15: Cumulative Posterior Probability of RS3129768 having an Association with SLE

Note: It appears from this plot that the MCMC algorithm has not fully converged to a posterior probability of RS3129768 having an association with SLE. However, over every repetition of running the algorithm, the posterior probability of this association is the same. The poor convergence is, again, likely to be due to the large number of

SNPs included in the model.

The posterior probabilities of each SNP are shown in the plot below.

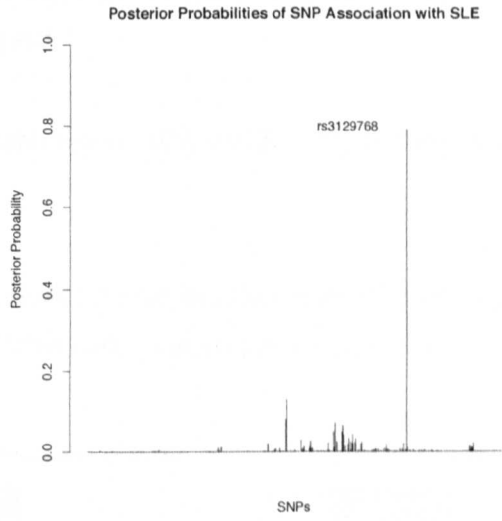


Figure 5.16: Posterior Probabilities by SNP of Association with SLE

This plot shows that SNP RS3129768 has the highest posterior probability of 0.79 of having an association with SLE. This SNP has a frequentist p-value of 2.1E-06 for an association with SLE, and was ranked 40th in the frequentist test. This SNP is not in the UK/US dataset. The highest signal in the frequentist analysis of the Spanish data has a posterior probability of 0.02 in the BMARS analysis. The top SNP, RS558702, from the UK/US BMARS analysis has a posterior probability of 0.02 in the Spanish analysis. The differences in the results could be due to the small sample size of the Spanish dataset and the lack of power.

Analysing the number of basis functions in the model, when RS3129768 is in the model, we found that there were was a probability of 0.0023 of only one basis function being in the model (a probability of 0.0003 of RS3129768 being in the model alone

and with no interaction term in the basis function), a probability of 0.66 of there being 2 in the model, and probabilities of 0.28 and 0.05 of 3 and 4 basis functions being in the model with RS3129768 respectively. From this we can conclude evidence of a second signal. This result is similar to that found in the UK/US analysis where there is a probability of 0.73 of there being 2 signals. There was a probability of 0 of no SNPs being included in the model.

There were prior probabilities of 0.08, 0.013, 0.002 of there being 2, 3 or 4 basis functions in the model.

Frequencies for knot values for each iteration of the BMARS algorithm including a basis function for RS3129768 were plotted to show the genotype's relationship with SLE.

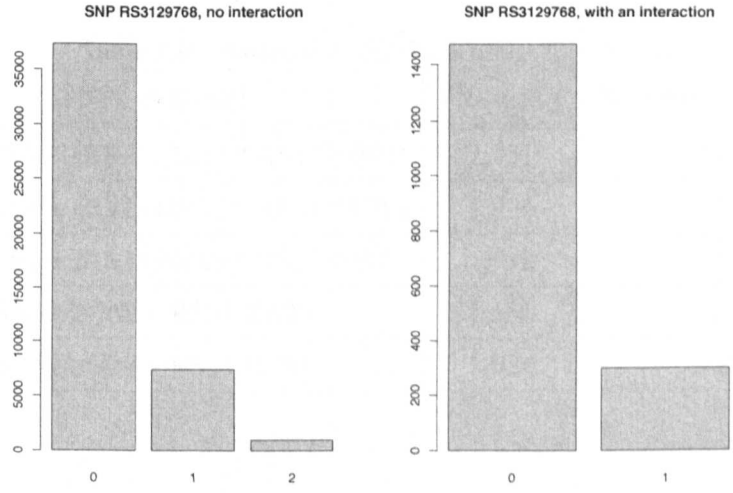


Figure 5.17: Frequencies of Knot Points for RS3129768

This shows that the relationship between RS3129768 and SLE is additive as a knot at 0 clearly has the highest posterior probability.

Investigating further into which SNPs appear together in the model with RS3129768, I found that the most common model is RS3129768(x_1) with another basis function with interaction between RS1793891(x_2) and RS3115663(x_3). This model has posterior probability of 0.05

$$\eta = -0.31 + 0.55([(x_1 - 0)]_+) + 0.60([-(x_2 - 2)]_+)([(x_3 - 0)]_+) \qquad (5.12)$$

Below is a table of posterior probabilities of the most common combination of SNPs. Note: These model frequencies are irrespective of knots.

Table 5.11: Posterior Probabilities of the Top Models

SNPs in Model	Frequency	Posterior Probability
RS3129768 + (RS3115663 * RS1793891)	2,751	0.05
RS3129768 + (RS2248902 * RS3130070)	1,936	0.03
RS3129768 + (RS3130626 * RS2248902)	1,918	0.03
RS3129768 + RS3130623	1,130	0.02
RS3129768 + RS3131381	1,024	0.02

Below are tables of LD measure D' and then r^2 of the top SNPs and those included in the most frequent models.

Table 5.12: D' Between Top SNPs

	RS3129768	RS3115663	RS1793891	RS2248902	RS3130070	RS3130626	RS3130623	RS3131381
RS3129768								
RS3115663	0.883							
RS1793891	0.284	0.232						
RS2248902	0.275	0.239	0.997					
RS3130070	0.882	1	0.235	0.241				
RS3130626	0.882	1	0.235	0.241	1			
RS3130623	0.886	0.996	0.241	0.247	1	1		
RS3131381	0.454	0.968	0.816	0.828	0.968	0.968	0.968	

Table 5.13: r^2 Between Top SNPs

	RS3129768	RS3115663	RS1793891	RS2248902	RS3130070	RS3130626	RS3130623	RS3131381
RS3129768								
RS3115663	0.039							
RS1793891	0.005	0.040						
RS2248902	0.005	0.043	0.987					
RS3130070	0.039	0.996	0.041	0.043				
RS3130626	0.039	0.996	0.041	0.043	1			
RS3130623	0.040	0.973	0.044	0.047	0.977	0.977		
RS3131381	0.004	0.338	0.1783	0.185	0.339	0.339	0.331	

These results, again, show evidence for 2 signals of association with SLE. The different SNPs in the top 3 models are in high LD with each other. RS3115663 is in LD with RS3130070 and RS3130626 with D' of 1. RS1793891 is in LD with RS2248902 with D' of 0.997. As with the UK/US analysis, the model has not selected between these 3 models with 2 basis functions which are in high LD. It could be possible that these top SNPs are due to a single underlying untagged locus.

For a plot of the posterior density of β for RS3129768 given it is in the model, please see Appendix 8.4

5.11.3 Imputed Dataset Using HapMap for Untyped SNPs

The BMARS algorithm described above was applied to 2,733 SNPs in this dataset for 3,723 individuals 100 times for 6 million iterations with a burn-in of 1 million and a thin of 1,000. Below is a convergence plot of the posterior probability of the highest signal in the model. For convergence plots of other top SNPs, please see Appendix 8.5.

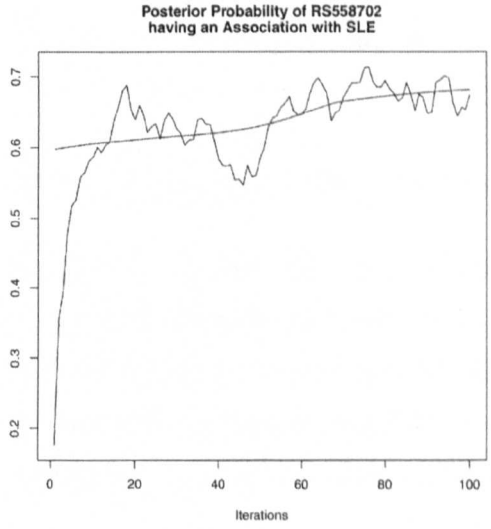


Figure 5.18: Convergence Plot of Posterior Probability of RS558702 having an Association with SLE

The posterior probabilities of each SNP having an association with SLE are shown below.

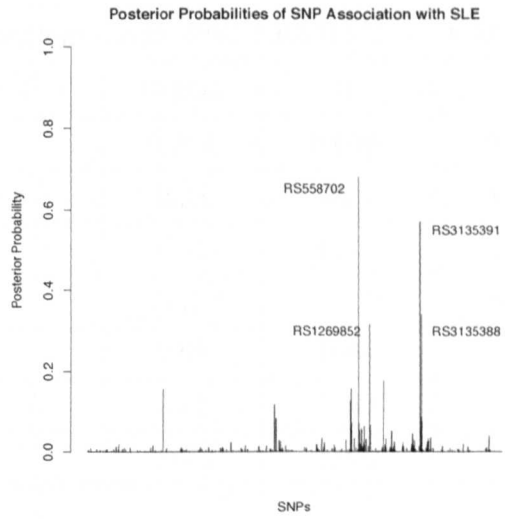


Figure 5.19: Posterior Probabilities by SNP of Association with SLE

This plot shows that yet again the strongest signal of an association with SLE is that from RS558702 with a posterior probability in this analysis of 0.68. The next highest signals (in order from left to right on the plot) are RS1269852, RS3135391 and RS3135388 with posterior probabilities of association 0.57, 0.34 and 0.32 respectively. The indicator variable for differences between UK/US and Spanish subjects had a posterior probability of 1.

The number of basis functions in the model when each of the top SNPs are in provides evidence for more than one signal. The table below shows the posterior probability of the number of basis functions in the model when each of the top SNPs is in the model.

Number of Basis Functions	RS558702	RS3135391	RS3135388	RS1269852	Overall
2	0.0004	0	0	0.0007	0.0005
3	0.012	0.006	0.008	0.013	0.012
4	0.25	0.25	0.27	0.27	0.26
5	0.41	0.42	0.42	0.42	0.41
6	0.23	0.23	0.22	0.22	0.23
7	0.08	0.08	0.07	0.07	0.07
8	0.02	0.015	0.01	0.01	0.01
9	0.002	0.002	0.002	0.002	0.002

Table 5.14: Posterior Probability of the Number of Basis Functions in the Model Given each of the Top SNPs are in the Model

This table shows that there is a posterior probability of 0.4 of there being 5 basis functions in the model. It is therefore likely that there are 4 signals in this data for an association with SLE. There is a posterior probability of 0.0005 of there only being one SNP in the model. There was a prior probability of 0.0002 of there being 5 basis functions in the model.

When SNP RS558702 is in the model there is a posterior probability of 0.01 of it being part of an interaction term. When SNPs RS3135391, RS3135388 and RS1269852 are in the model, there are a posterior probabilities of 0.04, 0.03 and 0.01 respectively of each SNP being part of an interaction term. There is a posterior probability of 0.99 that the indicator term for difference between UK/US and Spanish data is part of an interaction term in the model. The indicator term is interacting with 217 SNPs with posterior probabilities shown in the plot below.

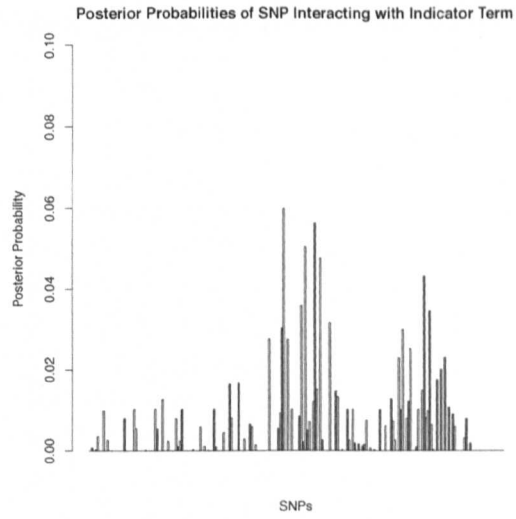


Figure 5.20: Posterior Probabilities by SNP of Interaction with Indicator Term

The inclusion of an interaction between a particular SNP with the indicator term implies that there is a difference of the effect of that SNP on SLE between studies. This difference is being accounted/ adjusted for by the inclusion of the indicator term. However, the posterior probabilities of the top SNPs being part of an interaction are very low. Therefore, I can conclude that there is no difference in the effect of the top SNPs on SLE between datasets.

The posterior probabilities of each SNP specifically interacting with the indicator term are all very low. Due to the high numbers of basis functions included in most of the models, the posterior probabilities of each specific basis function is thinly spread over a number of models. It is more important to consider the posterior probability of each SNP being in the model at all over all iterations.

Note: If the indicator term was included in the model independently, then this would imply that there is a difference between the frequency of cases and controls between the two datasets. In this case, there is a posterior probability of 0.30 of the indicator term being included in the model independently. This could be explained by the fact that 0.38 of individuals are cases in the UK/US dataset compared to 0.5 in the Spanish dataset.

Below is a table of the most common models and their posterior probabilities.

SNPs in Model	Frequency	Posterior Probability
Indicator + (RS3132550 * RS9268220) + RS558702 + (Indicator*RS1130288) + (RS3130288 * RS6906128) + RS3135391	1,626	0.003252
Indicator + (Indicator * RS410851) + RS558702 + (RS410851 *RS17839997) + RS3135391	1,571	0.003142
(Indicator * RS2858324) + (RS2844509*RS6457620) + RS558702 + RS3135391	1,522	0.003044
Indicator + (RS1966 *RS3094216) + RS558702 + (Indicator*RS7194) + RS7194 + (RS4947342 * RS2844510)	1,508	0.003016

The most common model (with specific knot points) has a posterior probability of 0.003. The effect on SLE on the probit scale is

$$\begin{aligned} y = & -0.42 + 0.04(I = Spanish) + 0.69[-(x_1 - 2)]_+ \\ & - 0.74[-(x_2 - 2)]_+ * (I = Spanish) \\ & + 0.26[(x_3 - 0)]_+ + 0.4035[-(x_2 - 2)]_+[-(x_4 - 1.99)]_+ \\ & - 0.36[-(x_5 - 2)]_+[(x_6 - 1.69)]_+ \end{aligned}$$

where x_1 represents RS558702, x_2 is RS3130288, x_3 is RS3135391, x_4 is RS6906128, x_5 is RS3132550 and x_6 is RS9268220. Note: these SNPs have posterior probabilities of an association with SLE of 0.67, 0.06, 0.57, 0.01, 0.08 and 0.02 respectively. As mentioned above, the SNPs spread are more thinly over several models. The posterior probabilities of individual SNPs over all models is more important.

Frequencies for knot values for each of the top SNPs given it is in the model are shown below to give an indication of their relationship with SLE.

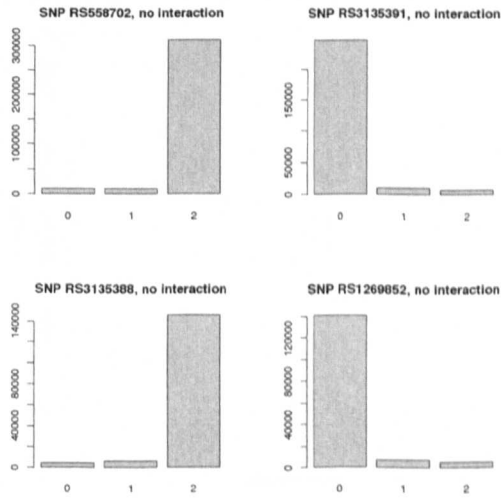


Figure 5.21: Frequencies of Knot Points for Top SNPs

From this is it evident that RS558702, RS3135391, RS3135388 and RS1269852 all have an additive effect on SLE.

The plots below show the posterior densities of β given that particular SNP is in the model

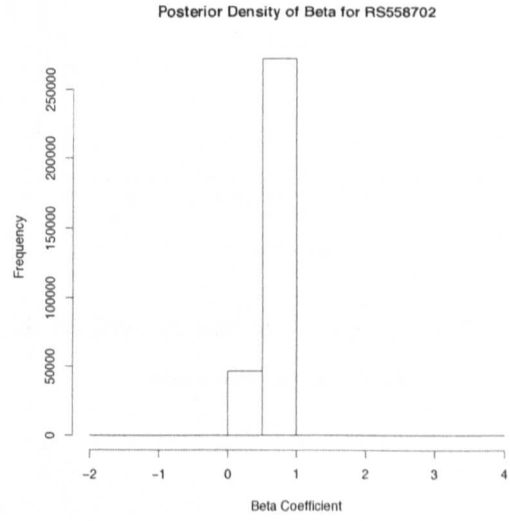


Figure 5.22: Posterior Density of β Associated with RS558702

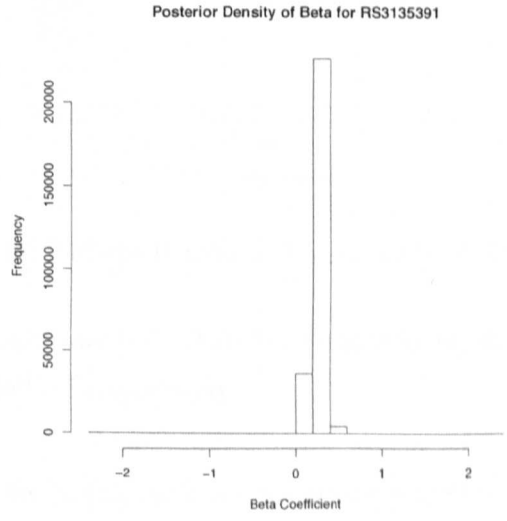


Figure 5.23: Posterior Density of β Associated with RS3135391

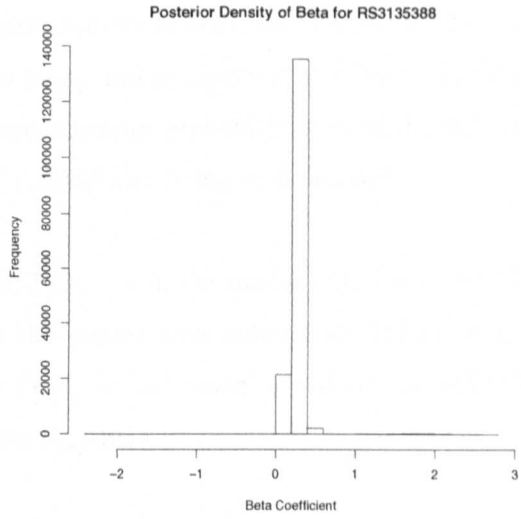


Figure 5.24: Posterior Density of β Associated with RS3135388

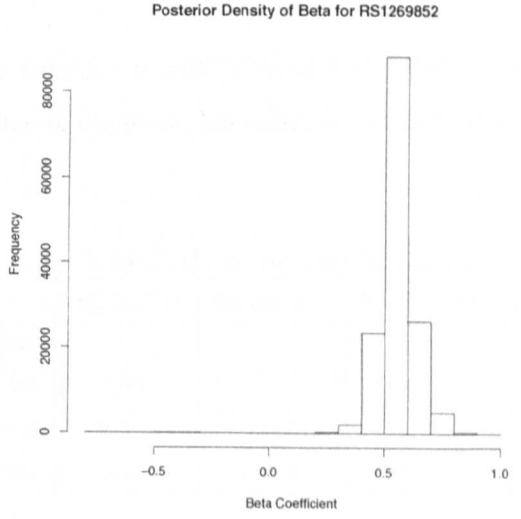


Figure 5.25: Posterior Density of β Associated with RS1269852

The mean beta coefficients are 0.57, 0.26, 0.27 and 0.55 for RS558702, RS3135391, RS3135388 and RS1269852 respectively.

When RS558702 is in the model, there is a posterior probability of 0.37 of RS3135391 also being in the model; 0.33 of RS3135388 also being in the model; and 0.0008 of RS1269852 also being in the model. Similarly, when RS3135391 is in the model, there are posterior probabilities of 0.68, 0.002, 0.31 of RS558702, RS3135388 or RS1269852 also being in the model respectively. When RS3135388 is in the model,

there are posterior probabilities of 0.67, 0.0008 and 0.32 of RS558702, RS3135391 and RS1269852 also being in the model respectively. Finally, when RS1269852 is in the model, there are posterior probabilities of 0.001, 0.54 and 0.33 of RS558702, RS3135391 and RS3135388 also being in the model.

This implies that RS558702 is in the model with either RS3135391 or RS3135388; or RS1269852 is in the model with either RS3135391 or RS3135388. RS558702 and RS1269852 are rarely in the model together; and RS3135391 and RS3135388 are rarely in the model together.

The tables below show LD measures D' and then r^2 for the top SNPs in the imputed dataset analysis.

The high LD values between RS558702 and RS1269852; and between RS3135391 and RS3135388 re-iterate the point that either is sufficient in the model in each case.

Table 5.15: D' Between Top SNPs

	RS558702	RS3135391	RS3135388	RS1269852
RS558702				
RS3135391	0.727			
RS3135388	0.722	1.000		
RS1269852	0.985	0.751	0.747	

Table 5.16: r^2 Between Top SNPs

	RS558702	RS3135391	RS3135388	RS1269852
RS558702				
RS3135391	0.014			
RS3135388	0.0138	0.996		
RS1269852	0.959	0.0148	0.0146	

5.12 Conclusions and Discusion

5.12.1 UK/US Analysis

It appears that there are 2 causal signals; most likely to be RS558702 and RS3135391. However, the top 4 SNPs are all in high LD so it is possible these signals are due to one underlying untagged locus.

5.12.2 Spanish Analysis

There is only one signal in this analysis, namely RS3129768. The top SNPs from the UK/US analysis have low posterior probabilities in this analysis. It is interesting that none of the other SNPs with small p-values from the frequentist analysis have high posterior probabilities. It is especially interesting that RS558702 or RS1269852 or any SNP in a similar position on the chromosome in high LD do not come up.

A simple frequentist generalised linear model to investigate the relationship between SLE and RS3129768 and the top 4 SNPs from the UK/US analysis, showed that the best fitting model was that with both RS558702 and RS3129768. In fact, the p-value for RS3129768 was more significant when conditioning on RS558702.

However, the convergence plots 5.15 show that the model has not converged properly. This could be due to the number of possible SNPs in the data to select from. The BMARS model can not cope with such a large dataset, especially with only 813 individuals.

In addition, the top SNP, RS3129768, from the Spanish dataset is only in LD with one other SNP in the SNPs with p-values $< 10^{-05}$ from the frequentist analysis. The top SNP from the UK/US and combined analyses, RS558702, on the other hand is in high LD (r^2 >0.8) with 25 other top SNPs from the frequentist analysis. This includes RS129852 which is second ranking SNP from both the UK/US and combined analyses. The other signal found in the UK/US analysis was either RS3135391 or RS3135388. These SNPs are in high LD with 4 other SNPs ranking highly from the frequentist

analysis.

The total posterior probabilities of all SNPs in LD (r^2 >0.8) with RS558702 is 0.51 i.e. there is a posterior probability of 0.51 that at least one of these SNPs is associated with SLE. This high posterior probability spread across several SNPs indicates that the model appears to be splitting the selection of particular signals between SNPs which are in high LD. It is possible that RS31269852 is selected as it is not in LD with any other SNPs that have a frequentist effect with SLE but it, itself does.

5.12.3 Analysis on Combined UK/US and Spanish Dataset

There are again, two main signals in this analysis. RS558702 with either RS3135391 or RS3135388; or RS1269852 with either RS3135391 or RS3135388. These are the same top 4 SNPs from the UK/US analysis. However, the top SNP from the Spanish analysis (RS3129768) is not included in the imputed dataset analysis as it was not in the UK/US dataset or in HapMap for imputation purposes.

If RS3129768 is removed from the Spanish BMARS analysis, then another signal (RS9271775) has a posterior probability of 0.8 and is again the only signal. RS9271775 is in high LD with RS3129768. This SNP, however, is in the imputed dataset but has a very low posterior probability, as shown in the table below.

Table 5.17: Table of Posterior Probabilities in Imputed Data Analysis of SNPs in High LD (D' of 1) with RS3129768 in Spanish Dataset

D'	r^2	Position	RS number	Posterior Probability
1	0.003	31647414	RS2516312	6.80E-05
1	0.018	32762692	RS9275184	0.00048
1	0.011	32674134	RS11754183	0.000178
1	0.019	32776824	RS9275383	0.000248
1	0.004	32911977	RS9378275	0.00015
1	0.004	32025270	RS2072634	0.000104
1	0.010	32038330	RS2734331	6.60E-05
1	0.018	32790115	RS3957148	0.000188
1	0.012	32904771	RS4148876	4.80E-05
1	0.009	31255073	RS9263823	0.000826
1	0.002	32232402	RS10947233	0.00016
1	0.009	32475975	RS3817964	0.000102
1	0.819	32702306	RS9271775	0.000152
1	0.029	32492505	RS9268541	0.001178
1	0.006	30046004	RS6457116	0.00012
1	0.017	32792235	RS9275614	8.40E-05
1	0.017	32793528	RS3916765	9.20E-05
1	0.027	31752619	RS13295	0.000328
1	0.029	32523953	RS13209234	0.000568
1	0.017	32789997	RS3998159	0.000358

RS9271775 is the best best predictor of RS3129786(with r^2=0.81 and D'=1) but it has a very low posterior probability of association with SLE in the imputed dataset analysis of 1.00E-04. From this, it is possible to conclude that given the higher power of the combined dataset and conditioning on RS558702, RS1269852, RS3135391 and RS3135388, the top SNP from the Spanish analysis does not have a high posterior probability of association with SLE. The Spanish dataset only has 813 individuals so it makes sense that these results, together with the conclusions above, on their own are not as convincing.

The results show that there are 5 basis functions in the model with posterior probability 0.42. This indicates that there are 5 signals in the data. It could be that there are more

underlying untyped causal SNPs along the chromosome. Rioux et al [46] concluded that there were two signals; RS1269852 and RS3135391. This study has shown that RS558702 has a higher posterior probability than RS1269852 but is also likely to be one of 2 signals with RS3135391. It has also highlighted that there may be 5 signals in the MHC region.

My BMARS model provides an automatic way of dealing with interactions and non-additive genotypic effects on SLE. BMARS does not assume a particular model and allows the data to select the SNPs with the highest posterior probabilities of association. It is a convenient and quick method for multivariate Bayesian model selection. It took less than 1 hour to run the BMARS analysis for 6 million iterations on a dataset with 2,733 SNPs for 3,723 individuals.

In a frequentist framework, tests of association could be carried out conditional on the top signals found in the SNP by SNP tests of association. However, there are more than 100 SNPs in the dataset with p-values of association $< 5 * 10^{-8}$. A step-wise regression model may lead to different results each time and not be time effective.

Under a Bayesian framework, uncertainty is quantified as the results give the probability a particular SNP has of being in the model.

In conclusion, I think BMARS is useful for datasets with $\leq$ 3,000 SNPs. (With more than 3,000 SNPs in high LD, the model has difficulty in selecting the true SNPs associated with SLE.) It selects the most important SNPs, gives a clear indication of the number of signals in the data and runs quickly. Finally, it is a relatively straight forward way of allowing for interactions and the possibility for different effects of each SNP.

CHAPTER

6

BAYESIAN NETWORKS FOR GENETIC ASSOCIATION STUDIES

6.1 Aims and Background

Genetic association studies have great potential to dissect the genetic basis of disease but raise a number of challenging and interesting statistical questions. These include both the potential complexity of the disease state, which may be categorised by a number of response variables, and the need to cope with large numbers of potential explanatory variables, both genetic and environmental. There are also complicated relationships between intermediate phenotypes or biomarkers; and genes and disease. Questions regarding how these are modelled jointly are important.

Coronary heart disease (CHD) is one of many diseases that results from complicated interactions between genetic and environmental factors. This means that identifying the genetic and environmental causal factors and understanding the relationship between disease and biomarker/intermediate traits is very difficult. By 1981, already

over 200 phenotypes had been shown to be associated with a higher risk of CHD [62]. I am interested in modelling the pathways between genotypes associated with CHD and a number of these phenotypes.

This study was particularly motivated by the work of Drenos et al [63]. They investigated the associations of genotypes with multiple blood biomarkers linked to CHD risk. It has been shown previously that the blood biomarkers associated with CHD are highly correlated with each other.

Changes in biomarkers such as lipid and lipoprotein particles and proteins involved in inflammation and coagulation, have a tendency to group amongst those patients with a higher risk of CHD. This makes it difficult to determine the relationships and direction of these relationships with CHD. Due to the correlation amongst these biomarkers, establishing an independent effect on CHD outcome is hard. The associations found between the blood phenotypes and CHD could be causal. However, it could be that reverse causation or confounding is present.

It has been shown that several candidate gene and genome wide association studies [64–66] that nearly all these highly correlated biomarkers associated with CHD are also associated with SNPs. Drenos et al examined several SNPs associated with CHD, and found that these SNPs also had relationships with several blood phenotypes. From their findings, Drenos et al propose that information on genotype and blood phenotypes may be used to disentangle the complicated relationships with disease outcome.

As discussed in 4.8, due to Mendelian randomisation, genotype allocation is considered to be randomly assorted. It is therefore possible to use these genotypes as possible instrumental variables in analysing the complicated associations and directions of relationships between these biomarkers. Consequently, associations between CHD and genotype; and blood phenotype and genotype should not be subject to the problems of confounding because genotypes do not vary due to phenotype, disease or any exogenous factors.

The analysis done by Drenos et al is limited in that only univariate associations be-

tween the genotypes, phenotypes and disease outcome are highlighted. I aim to develop methods for the joint analysis several phenotypes and genotypes using directed acyclic graphs (DAGs). Joint analyses of all this information could be used to disentangle CHD pathways amenable for possible treatment. DAGs allow us to model data jointly and therefore correct for other variables within the model. This allows us to get a better understanding of the underlying structure of the data and possibly untangle correlated factors. I aim to use these methods for modelling intermediate phenotypes and disease outcomes to better understand the joint relationships between them.

The relationships I will focus on are between APOE, CETP and APOB genotypes; and HDL- and LDL- cholesterol, triglycerides, C-reactive protein, and apolipoproteins A and B after these were highlighted by Drenos et al.

6.2 NPHS-II Data

The Northwick Park Heart Study II (NPHS-II) on coronary heart diseases is a prospective study of 3012 healthy middle-aged men aged 50-64 years at recruitment, sampled from nine UK general practices between 1989 and 1994. Measures were made of at least 15 circulating blood factors associated with CHD risk that included both circulating proteins, and non-protein metabolites. By December 2005, after a median follow-up of 13.6 years, there had been 296 definite fatal or non-fatal CHD events (230 in 2401 of the genotyped sample. [63]

The table and plot below summarise the relationships found by Drenos et al.

	BMI	Systolic	Diastol	Chol	HDL	LDL	Triglyc	ApoB	ApoAI	Homoc	Folate	LpPLA2	CRP	FVII
Systolic BP	0.207**													
Diastolic BP	0.248**	0.707**												
Cholesterol	0.088**	0.100**	0.082**											
HDL	-0.186**	-0.096**	-0.166**	-0.014										
LDL	0.067**	0.032	0.016	0.794**	-0.140**									
Triglycerides	0.326**	0.225**	0.233**	0.318**	-0.489**	0.081**								
ApoB	0.147**	0.080**	0.030	0.601**	-0.145**	0.598**	0.255**							
ApoAI	-0.140**	-0.003	-0.040	0.101**	0.517**	0.038	-0.207**	-0.243**						
Homocyst	-0.047	0.092**	0.047	-0.019	-0.011	-0.016	0.013	0.033	-0.037					
Folate	0.110**	0.029	0.012	0.037	-0.007	-0.004	0.108**	0.010	0.029	-0.463**				
LpPLA2	0.080**	0.099**	0.095**	0.314**	-0.245**	0.281**	0.231**	0.193**	-0.080**	0.046	0.015			
CRP	0.251**	0.175**	0.121**	0.103**	-0.212**	0.079**	0.230**	0.134**	-0.171**	0.052	0.043	0.043*		
FVII	0.082**	0.096**	0.120**	0.226**	-0.029	0.116**	0.257**	0.176**	0.024	0.065*	-0.013	0.006	0.128**	
FIB	0.072**	0.123**	0.062*	0.095**	-0.182**	0.121**	0.068**	0.153**	-0.154**	0.061*	-0.037	0.018	0.434**	0.094

Pearson's correlation coefficients

1.000-0.4000

0.399- 0.200

0.199-0.080

0.079-0.030

0.029- 0.020

0.019-0.004

Figure 1. Correlations between multiple phenotypes linked to CHD in 2775 men from the NPHSII study. Values in cells indicate Pearson's correlation coefficient R. $^{*}P < 0.01$, $^{**}P < 0.001$ (see colour code). Baseline and five repeat measures were available for cholesterol, triglycerides (TG), coagulation factor VII (FVIIc), fibrinogen, blood pressure (BP), smoking and body mass index (BMI) and single measures for the remaining traits.

Figure 6.1: Correlation between phenotypes of men from NPHS [63]

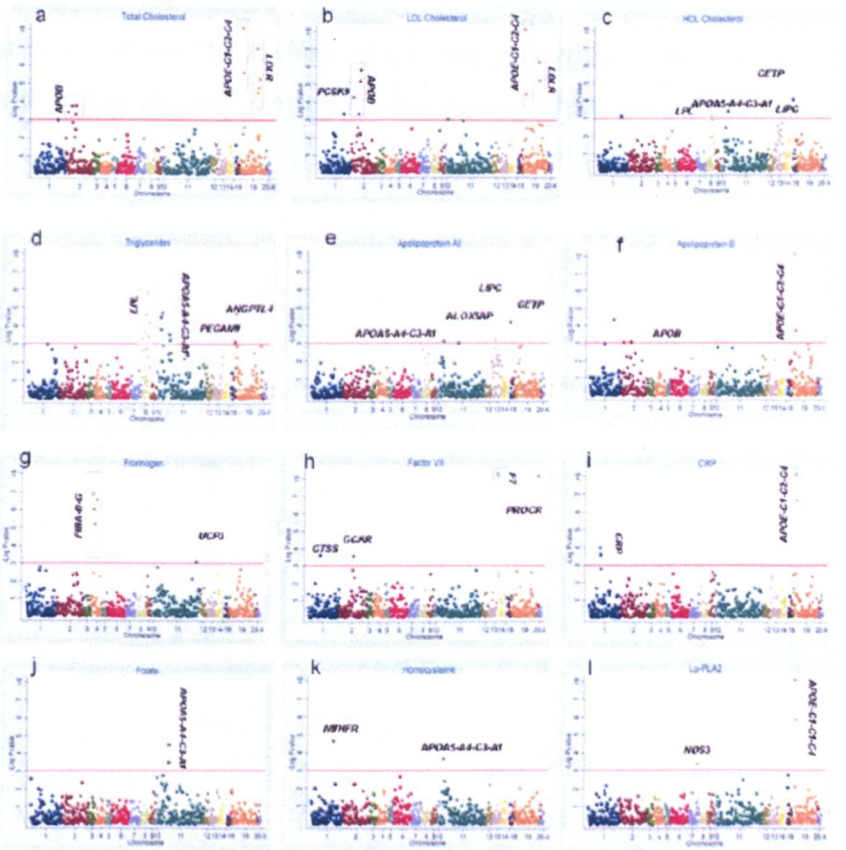


Figure 2. Associations of 860 SNPs by chromosome with 12 blood phenotypes. The horizontal line indicates a critical FDR threshold of 0.2, approximately equivalent to a P-value $< 10^{-3}$.

Figure 6.2: Association between SNPs and phenotypes of men from NPHS [63]

Figure 6.1 shows the relationships between phenotype measures and it is apparent that they are correlated with each other. Figure 6.2 highlights the relationships between SNPs and phenotypes of men in the NPHS. It is obvious that the phenotypes are associated with more than one genotype, and several are associated with the same genotypes. For example, APOE genotype is associated with both CRP level and APOB level. These relationships, and other subsets, will be discussed later in Section 6.4.

I will investigate how these relationships work jointly. Much is known about biological associations for coronary heart disease between certain genes and phenotypes, and so certain results can be expected. If these results are not found then it may be an indicator that the proposed algorithm is ineffective of selecting the "correct" model, and so these data are perfect for initial testing purposes as well as further explanation.

The plot below shows the number of missing genotypes in my data by individual. Those with >30% missing genotypes were deleted which left a dataset of 2,385 individuals.

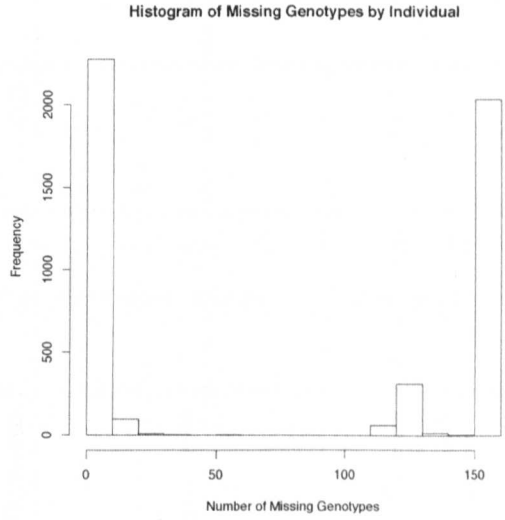


Figure 6.3: Histogram of number of missing genotypes by individual

Once those with >30% missing genotypes were deleted, the number of missing individuals by genotypes were as described in the table below. The table gives a summary of the the variables used in my PhD from the NPHS-II.

PHENOTYPES	Description	Distribution	Mean	SD	% Missing	Any known biological associations
HDL level	High-density lipoprotein (HDL) cholesterol enables lipids like triglycerides to be transported within the bloodstream	Gaussian	1.7	0.59	1.38	Genotypes: CETP, LPL, APOA, Phenotypes: LDL level, TG level, APOB level, APOA level, CRP level & correlated with TG level, APOA level
TG level	Triglycerides (TG) play an important role in metabolism as energy sources and transporters of dietary fat.	Log Gaussian	0.58	0.53	0.55	Genotypes: LPL, APOA, Phenotypes: HDL level, LDL level, APOB level, APOA level, CRP level correlated with HDL level
Smoking	Smoker or non-smoker?	1700 non-smokers, 685 smokers			0.00	
CRP level	C-reactive protein (CRP) is a protein found in the blood, and is synthesised by the liver in response to factors released by fat cells.	Log Gaussian	1.1	1.17	8.55	Genotypes: CRP, APOE, Phenotypes: HDL level, LDL level, TG level, APOB level, APOA level
APOB level	Apolipoprotein B (APOB) is the primary apolipoprotein of low-density lipoproteins (LDL), which is responsible for carrying cholesterol to tissues	Log Gaussian	-0.14	0.28	13.17	Genotypes: APOB, APOE, Phenotypes: HDL level, LDL level, APOA level
GENOTYPES		0s	1s	2s		
LPL	Encodes for Lipoprotein lipase (LPL) which interacts with TG	46	630	1709	0.00	Phenotypes: HDL level, TG level
CETP	Encodes for Cholesteryl ester transfer protein (CETP) which transports triglycerides between the lipoproteins	693	1288	404	7.00	Phenotypes: HDL level, APOA level
	Encodes for Apolipoprotein A (APOA) which has a role in lipid metabolism	2068	302	15	0.13	Phenotypes: HDL level, TG level, APOA level
APOE	Encodes for Apolipoprotein E (APOE) which has a role in lipid metabolism	1390	414	581	1.13	Phenotypes: LDL level,APOB level, CRP level
CRP	Encodes for CRP	1053	1072	260	1.13	Phenotypes: CRP level
APOB	Encodes for APOB	1953	403	29	6.33	Phenotypes: LDL level, APOB level

APOE is defined as a two SNP variant: rs429358 + rs7412. It is known that these two SNPs work biologically together. [67] Together they are commonly known as the APOE 'genotype', and is how they will be referred to from now in this PhD. For the other genotypes in my analysis, I used one SNP per gene and these are summarised below.

Genotype	RS number
LPL	rs264
CETP	rs708272
APOA	rs6589566
CRP	rs3091244
APOE	rs429358+rs7412
APOB	rs585967

The tables below show the LD measures between the genotypes used in my analysis; firstly D' and then r^2. It is clear that these genotypes are not correlated with each other.

	LPL	CETP	APOA	CRP	APOE	APOB
LPL						
CETP	0.016					
APOA	0.174	0.058				
CRP	0.016	0.081	0.277			
APOE	0.0371	0.066	0.300	0.019		
APOB	0.030	0.051	0.014	0.111	0.105	

	LPL	CETP	APOA	CRP	APOE	APOB
LPL						
CETP	3.45e-05					
APOA	4.02e-04	1.98e-04				
CRP	8.55e-05	3.07e-04	3.41e-04			
APOE	1.04e-03	1.31e-03	1.59e-03	9.53e-05		
APOB	2.69e-05	3.56e-04	8.60e-05	1.28e-04	4.51e-04	

6.3 Methods

I expanded methodology proposed by Fronk and Giudici [68] for Markov Chain Monte Carlo (MCMC) selection for directed acyclic graphs (DAGs). Fronk and Giudici propose methods to cycle over possible DAGs searching for the relationships with the highest posterior probabilities between nodes, using an MCMC with reversible jump, developed by Green [34]. I expand their methodology to a genetic context, and to allow for specific issues relating to modelling complex disease pathways between genotypes, intermediate phenotypes and disease.

6.3.1 Directed Acyclic Graphs

In this framework, the DAG (see Chapter 4 for clarification) is set up so that each node is a variable; a genotype or phenotype. An arrow between them indicates a direct association between them, which or or may not be informative about the direction of that association. The joint model of all data will automatically correct for the effects of all variables in the model via the edges defined. This model allows us to distinguish between direct and indirect effects as well as explore possible directionality of relationships. Since different DAGs can belong to the same equivalence class (see Section 4.5), directions of association may become indistinguishable and I am interested in the implications of this. Note that in this model a genotype can not be dependent on a phenotype as this does not make biological sense. A genotype can not be determined by levels of phenotype.

The joint model of all the data will automatically correct for all the effects of all variables included in the model via the edges defined. Therefore if I allow the algorithm to choose the most appropriate model any confounders included in the model will be automatically corrected for.

6.3.2 Algorithm Overview

A brief overview of the algorithm proposed is outlined below. This is then explained in more detail.

- Start at any (legal) DAG model.
- Propose a directed arrow between two variables.
- When an association is already included in the model, check that addition or reversal of proposed arrow to current DAG does not lead to a cyclic graph, and that direction of arrow makes sense. Note: As discussed above, a genotype can not be dependent on a phenotype.
- Reversible jump MCMC; select birth, death or switch step at each iteration dependent on current DAG and directed arrow proposed.
- Accept or reject proposed change in arrow between two variables.
- Gibbs sampler, based upon current DAG, to update model parameters.
- Iterate until convergence.
- Analyse DAGs with highest posterior probabilities.

6.3.3 Bayesian Multivariate Gaussian Linear Regression

Given the current DAG, d, and following methodology developed by Fronk & Giudici [68], I jointly consider a Gaussian regression model for each phenotype X_i regressed upon it's parents (genotype or phenotype) for i=1,...,p, where p is the number of nodes in the model. The number of parents, pa(i), varies by node and the set of

parents for node X_i is denoted by matrix $\boldsymbol{x}_{pa(i)}$ with dimension n x p_i. n is the number of individuals in the dataset. Each regression model is defined as:

$$X_i|\boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \boldsymbol{\sigma}_i^2, d \sim N(\beta_{io} + \sum_{x_l \in pa(x_i)} \beta_{il} x_l, \sigma^2_{i|pa(i)})$$

where β_{io} is the intercept term, β_{il} are the regression coefficients, pa(i) indicates the parents of node X_i and σ_i^2 is the partial variance of X_i given parents pa(i).

The priors for β, σ and d are given by:

$$\beta_{i|pa(i)}|\sigma^2_{i|pa(i)}, d \sim N(\boldsymbol{b}_{i|pa(i)}, \sigma^2_{i|pa(i)}\boldsymbol{I})$$

$$\sigma^2_{i|pa(i)}|d \sim IG(\delta_{i|pa(i)}, \lambda_{i|pa(i)})$$

p(d) = $\frac{1}{D}$

where D is the total number of possible DAGs, given the number of nodes X_i. This is discussed more below in Section 6.3.5. I is the identity matrix with a dimension of the number of parents + 1 for the intercept term. This assures the coefficients of this regression model to be mutually independent, a priori. σ is given an inverse gamma prior for computational simplicity as this is the conjugate (see Section 3.3) for the variance of a Gaussian prior.

Again, as in Fronk & Giudici, using the factorisation and global parameter independence properties of joint distributions (for details see [69]), the joint distribution is given by

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d) &= p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, d)p(\boldsymbol{\beta}|\boldsymbol{\sigma}^2, d)p(\boldsymbol{\sigma}^2|d)p(d) \\
&= \prod_{i=1}^{p} p(x_i|\boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}) \prod_{i=1}^{p} p(\boldsymbol{\beta}_{i|pa(i)}|\sigma^2_{i|pa(i)}) \prod_{i=1}^{p} p(\sigma^2_{i|pa(i)})p(d) \\
&= \prod_{i=1}^{p} (2\pi\sigma^2_{i|pa(i)})^{\frac{-n}{2}} exp(-\frac{1}{2\sigma^2_{i|pa(i)}} \sum_{l=1}^{n} (x_{li} - \beta_{i0} - \boldsymbol{\beta}_{i|pa(i)}\boldsymbol{x}_{pa(i)})^2) \\
&\quad * \prod_{i=1}^{p} (2\pi\frac{1}{\alpha}\sigma^2_{i|pa(i)})^{-\frac{1}{2}} exp(-\frac{\alpha}{2\sigma^2_{i|pa(i)}} (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)})^T (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)})) \\
&\quad * \prod_{i=1}^{p} \frac{\lambda^{\delta_{i|pa(i)}}_{i|pa(i)}}{\Gamma(\delta_{i|pa(i)})} (\sigma^2_{i|pa(i)})^{\delta_{i|pa(i)}-1} exp(-\frac{\delta_{i|pa(i)}}{\sigma^2_{i|pa(i)}}) * \frac{1}{D}
\end{aligned}
\tag{6.1}
$$

6.3.4 Matrix of Allowed Direction

As mentioned previously, it is not possible for a genotype to be a child as it cannot be determined by phenotype level or disease status. It is also assumed for this analysis in particular that the genotypes do not have any association with any other genotype as they are not in LD. Therefore, the DAG model search space needs to be limited in my algorithm to account for this. i.e. there can never be an arrow from a phenotype node into a genotype one, or any arrows between genotypes.

After proposing an arrow between two nodes at each iteration, it is checked that the proposed association and dependency makes sense. If not, then another arrow will be proposed until it does. This is done using a p*p matrix of 0s and 1s (where p is the number of nodes) was developed. The ith row and the jth column is 0 if a move from parent i to child j is not allowed, and 1 otherwise. The 0s and 1s are dependent on whether the ith row and jth column indicate a directed association between a genotype and phenotype, or between phenotypes.

6.3.5 Test for Acyclicity

At each iteration, the proposition of a new arrow must also be tested to check whether its addition to the current DAG model will create a cycle.

A graph is acyclic if it contains no cycles. i.e. there are no closed loops.

The proposed DAG will not be acyclic if and only if one of the following steps is true

1. one of the parents of the new proposed parent is the child proposed
2. one of the grand-parents of the new proposed parent is the child proposed
3. one of the great grand-parents of the new proposed parent is the child proposed
4. etc. depending on how many nodes could be included in a possible loop. i.e. there is a maximum of (p-1) possible generations

If the arrow fails this check, then another is proposed until I am proposing a new directed acyclic graph.

Prior on DAGs

Under the framework used by Fronk and Giudici, the prior on each DAG, d, is $\frac{1}{D}$ where D is the total number of possible DAGs, given the number of possible nodes X_i in the model. In this case, each possible DAG is selected with equal probability.

A new arrow is proposed randomly with equal probability. Fronk and Giudici have no restrictions on the relationships on the relationships between nodes: each possible DAG with the addition of a new arrow proposed is allowed. However, given the restricted direction of association from a genotype to a phenotype in my proposed model, it is not clear whether each possible DAG has an equal probability a priori of being selected. Given the prior of equal probability for every possible arrow given restrictions, I investigated whether the model would favour 'mid-size' DAGs.

In order to investigate if this would favour any DAG model in particular, I ran a MCMC algorithm for 6 nodes with no likelihood function, data or priors on other model parameters. The model used was to propose iteratively an arrow between two nodes

uniformly at random.

All proposed steps were accepted with probability 1. The results gave probabilities of selecting each possible DAG and numbers of arrows included at each iteration. Below is a histogram of the number of arrows in the model over 200,000 iterations.

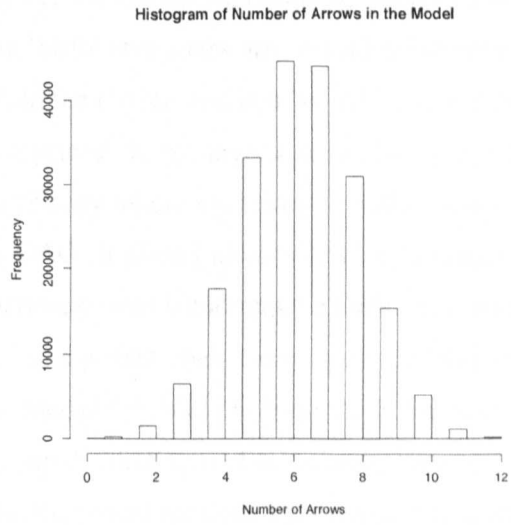


Figure 6.4: Histogram of number of arrows in D under prior of uniformly selected arrows

This shows that under the prior of equal weight to each possible arrow at each iteration, there is a prior probability of 0.17 of 5 arrows, 0.22 of 6 arrows and 0.22 of 7 arrows out of a maximum of 12 in this particular model with 6 nodes and 3 genotypes. Therefore, this prior favours 'mid-size' DAGs. This will be taken into consideration in analysis of the results.

6.3.6 Application Using Reversible Jump MCMC

I wish to estimate posterior probability $p(d, \boldsymbol{\beta}, \boldsymbol{\sigma^2}|\boldsymbol{x})$ and the marginal distributions $p(d|\boldsymbol{x})$, $p(\boldsymbol{\beta}|\boldsymbol{x})$ and $p(\boldsymbol{\sigma^2}|\boldsymbol{x})$.

Again, following the methodology outlined by Fronk and Giudici, I use reversible jump MCMC to sample from the joint posterior distribution of d, β and σ in order to make inference about these parameters; in particular to discover the most likely models underlying the data (i.e. d with highest posterior probability). My aim is to cycle over possible DAGs to find those with the highest posterior probabilities. As described in Section 3.7, a reversible jump algorithm allows sampling over models with varying dimension, as is the case with my DAG model space.

To re-iterate, at each iteration, arrows are randomly proposed to be added, dropped or direction swapped. In a 'birth' step a new arrow is added to the model; in a 'death' step an arrow is dropped from the model; and in a 'switch' step the direction of association between two nodes is reversed. In my algorithm, a child j, and parent i, are randomly selected, testing for acyclicity where necessary to make sure such a move would not produce a non-acyclic DAG. If child j already has i as a parent then it is proposed to drop the association between i and j, and remove node i as a parent of j (a death step). If child j doesn't have i as a parent, then it is proposed to add an association from i to j, and add node i to the model for child j (a birth step). However, if i has j as a parent, then it is proposed to swap the direction of association between the two nodes so that i becomes the parent, and the model for node i gains a parent, whilst the model for node j loses a parent.

If a birth step is proposed, then a new coefficient for the association of i being a parent of j is introduced, β'_{ij}. The proposal distribution of the new coefficient is $q(\beta'_{ij}) \sim N(0, \eta^2)$. The acceptance probability of this step is reduced to:

$$\text{prob accept} = min\left(1, \frac{p(x_j|\boldsymbol{x}_{pa'(j)}, \boldsymbol{\beta}_{j|pa'(j)}, \sigma^2_{j|pa'(j)})p(\boldsymbol{\beta}_{j|pa'(j)}|\sigma^2_{j|pa'(j)})}{q(\beta'_{ji})p(x_j|\boldsymbol{x}_{pa(j)}, \boldsymbol{\beta}_{j|pa(j)}, \sigma^2_{j|pa(j)})p(\boldsymbol{\beta}_{j|pa(j)}|\sigma^2_{j|pa(j)})}\right) \quad (6.2)$$

where $\boldsymbol{x}_{pa'(j)}$ refers to the proposed (new) matrix of parents j for child i i.e. this matrix has increased in dimension by 1, $\boldsymbol{\beta}_{j|pa'(j)}$ refers to the proposed (new) matrix of βs which now includes another β for the extra parent of i, and similarly for $\sigma^2_{j|pa'(j)}$. $\sigma^2_{j|pa'(j)} = \sigma^2_{j|pa(j)}$ and p(d) is uniform so there is no need to include the $\frac{p(\sigma^2_{j|pa'(j)}|d)p(d)}{p(\sigma^2_{j|pa(j)}|d)p(d)}$

part of the acceptance ratio as it cancels out.

If the step is not accepted then the DAG remains the same for this iteration. If a death step is proposed, then the acceptance probability is essentially the reciprocal of that for the birth step, and I propose to drop the coefficient β_{ij} relating to the directed association from i to j.

If a switch step is proposed then, i loses j as a parent, but j gains i as a parent. Therefore, I lose the regression coefficient β_{ji} from $\beta_{i|pa(i)}$, but I gain β'_{ij} to the vector $\beta_{j|pa(j)}$. The proposal distributions, then, are those proposed by Gelman (1995):

$$q_\sigma = \sigma'^2_{i|pa'(i)} | \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)} \sim \text{Inv-}\chi^2(n - |pa'(i)|; s'^2) \tag{6.3}$$

where $s'^2 = \frac{1}{n-|pa'(i)|}(\boldsymbol{x}_i - \boldsymbol{X}_{pa'(i)}\hat{\beta}')^T(\boldsymbol{x}_i - \boldsymbol{X}_{pa'(i)}\hat{\beta}')$, and $\hat{\beta}' = (\boldsymbol{X}^T_{pa'(i)}\boldsymbol{X}_{pa'(i)})^{-1}\boldsymbol{X}^T_{pa'(i)}\boldsymbol{x}_i$

$$q_\beta = \beta_{i|pa'(i)} | \sigma'^2_{i|pa'(i)}, \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)} \sim N(\hat{\beta}', \sigma'^2_{i|pa'(i)} V') \tag{6.4}$$

where $V' = (\boldsymbol{X}'_{pa'(i)}\boldsymbol{X}_{pa'(i)})^{-1}$. The proposal distributions for $\beta'_{j|pa'(j)}$ and $\sigma'^2_{j|pa'(j)}$ are derived similarly. Fronk and Giudici suggest that these proposal distributions are used to achieve high acceptance rates by proposing to assign new values for all parameters associated with i and j at a specific iteration. The parameters are drawn from the current model. The acceptance probability of this step is:

$$\begin{aligned}
\text{prob accept} \;=\; & min(1, \frac{p(x_j|\boldsymbol{x}_{pa'(j)}, \beta'_{j|pa'(j)}, \sigma'^2_{j|pa'(j)})p(\beta'_{j|pa'(j)}|\sigma'^2_{j|pa'(j)})p(\sigma'^2 j|pa'(j))}{p(x_j|\boldsymbol{x}_{pa(j)}, \beta_{j|pa(j)}, \sigma^2_{j|pa(j)})p(\beta_{j|pa(j)}|\sigma^2_{j|pa(j)})p(\sigma^2_{j|pa'(j)})} \\
& * \frac{p(x_i|\boldsymbol{x}_{pa'(i)}, \beta'_{i|pa'(i)}, \sigma'^2_{i|pa'(i)})p(\beta'_{i|pa'(i)}|\sigma'^2_{i|pa'(i)})p(\sigma'^2 i|pa'(i))}{p(x_i|\boldsymbol{x}_{pa(i)}, \beta_{i|pa(i)}, \sigma^2_{i|pa(i)})p(\beta_{i|pa(i)}|\sigma^2_{i|pa(i)})p(\sigma^2_{i|pa'(i)})} \\
& * \frac{q_\beta(\beta_{j|pa(j)}|\sigma^2_{j|pa(j)}, \boldsymbol{x}_j, \boldsymbol{X}_{pa(j)})q_\sigma(\sigma^2_{j|pa(j)}|\boldsymbol{x}_j, \boldsymbol{X}_{pa(j)})}{q_\beta(\beta'_{j|pa'(j)}|\sigma'^2_{j|pa'(j)}, \boldsymbol{x}_j, \boldsymbol{X}_{pa'(j)})q_\sigma(\sigma'^2_{j|pa'(j)}|\boldsymbol{x}_j, \boldsymbol{X}_{pa'(j)})} \\
& * \frac{q_\beta(\beta_{i|pa(i)}|\sigma^2_{i|pa(i)}, \boldsymbol{x}_i, \boldsymbol{X}_{pa(i)})q_\sigma(\sigma^2_{i|pa(i)}|\boldsymbol{x}_i, \boldsymbol{X}_{pa(i)})}{q_\beta(\beta'_{i|pa'(i)}|\sigma'^2_{i|pa'(i)}, \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)})q_\sigma(\sigma'^2_{i|pa'(i)}|\boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)})})
\end{aligned} \tag{6.5}$$

where the 's indicate matrices or vectors with dimensions of model proposing. For example, $x_{pa'(i)}$ relates to the matrix of parents of i, including a column for proposed child j as a result of the proposed switch step.

The final step of the algorithm is to update the regression coefficients, and their variance with the new vectors $\beta'_{i|pa(i)}$ and $\sigma^2_{i|pa(i)}$ using the Gibbs sampler based upon their full conditional distributions:

$$\beta'_{i|pa(i)} \sim N(m_{i_{full}}, \sigma_{i_{full}}) \tag{6.6}$$

where $m_{i_{full}} = \sigma_{i_{full}}(\frac{1}{\sigma^2_{i|pa(i)}} \boldsymbol{X}^T_{pa(i)} \boldsymbol{x}_i + \frac{\alpha}{\sigma^2_{i|pa(i)}} \boldsymbol{b}_i)$ and $\sigma_{i_{full}} = \sigma^2_{i|pa(i)}(\boldsymbol{X}^T_{pa(i)} \boldsymbol{X}_{pa(i)} + \alpha \boldsymbol{I})^{-1}$

$$\sigma^2_{i|pa(i)} \sim IG(\delta_{i_{full}}, \lambda_{i_{full}}) \tag{6.7}$$

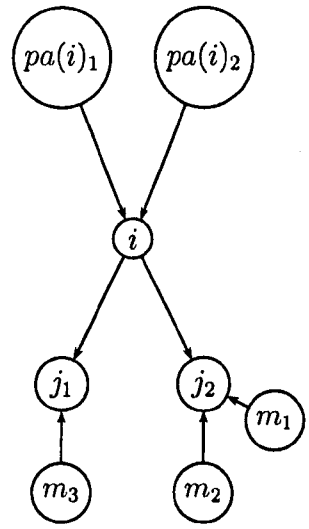
where $\delta_{i_{full}} = \delta_{i|pa(i)} + 0.5(n+p)$ and
$\lambda_{i_{full}} = \lambda_{i|pa(i)} + 0.5((\boldsymbol{x}_i - \boldsymbol{X}_{pa(i)}\boldsymbol{\beta}_{i|pa(i)})^T(\boldsymbol{x}_i - \boldsymbol{X}_{pa(i)}\boldsymbol{\beta}_{i|pa(i)}) + (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)})^T(\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)}))$

6.3.7 Missing Data

Missing genotypes were imputed using Mach [17]. I ran the MACH program for 50 iterations, considering 200 haplotypes at each iteration . This was reasonable for the small amount of missing data, as shown in the histogram above 6.3. Each of these had a minimum r^2 of 0.9. r^2 estimates the squared correlation between imputed and true genotypes. I used the expected genotype value for each missing genotype.

I assume the missing data to be missing at random (MAR) [70]. The missing phenotype data was imputed using a Gibbs sampler within the above algorithm. At each iteration, the missing data was sampled using the conditional distribution of the missing data at a specific node given current values of all parents, children and related parameters of the current DAG. Consider the example of a DAG below to illustrate the joint model

to be considered when sampling the missing values of $\boldsymbol{x}$. i has two parents $pa(i)_1$ and $pa(i)_2$, two children j_1 and j_2 with parents other than i m_1 and m_2; and m_3 respectively.



Denoting $x_{i_{miss}}$ as the vector of missing phenotypes of x_i for each node i= 1, ..., p, the number of nodes in the DAG at each particular iteration, and the notation used above, the conditional distribution is given by

$$
\begin{aligned}
p(\boldsymbol{x}_{i_{miss}}|\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\sigma}^2,d) &\propto p(\boldsymbol{x}_i|\text{parents}(\boldsymbol{x}_i)) \prod_{\text{j s.t. j is a child of i}} p(\boldsymbol{x}_j|\text{j is a child of i}) \\
&\propto p(\boldsymbol{x}_{i_{miss}}|\boldsymbol{x}_{pa(i)_{miss}},\boldsymbol{\beta}_{i|pa(i)},\sigma^2_{i|pa(i)},d) * \prod_{\text{j s.t. j is a child of i}} p(\boldsymbol{x}_{j_{miss}}|\boldsymbol{x}_{pa(j)miss(i)},\boldsymbol{\beta}_{j|pa(j)},\sigma^2_{j|pa(j)},d) \\
&= (2\pi\sigma^2_{i|pa(i)})^{\frac{-n_{imiss}}{2}} exp(-\frac{1}{2\sigma^2_{i|pa(i)}}\sum_{l}^{n_{imiss}}(x_{li_{miss}}-\beta_{i0}-\beta_{i|pa(i)}\boldsymbol{x}_{pa(i)_{miss}})^2) \\
&* \prod_{\text{j s.t. j is a child of i}} (2\pi\sigma^2_{j|pa(j)})^{\frac{-n_{imiss}}{2}} exp(-\frac{1}{2\sigma^2_{j|pa(j)}}\sum_{l}^{n_{imiss}}(x_{lj_{miss(i)}}-\beta_{j0}-\beta_{j|pa(j)}\boldsymbol{x}_{pa(j)_{miss(i)}})^2) \\
&\propto exp(-\frac{1}{2\sigma^2_{i|pa(i)}}\sum_{l}^{n_{imiss}}(x_{li_{miss}}-\beta_{i0}-\beta_{i|pa(i)}\boldsymbol{x}_{pa(i)_{miss}})^2) \\
&* \prod_{\text{j s.t. j is a child of i}} exp(-\frac{1}{2\sigma^2_{j|pa(j)}}\sum_{l}^{n_{imiss}}(x_{lj_{miss(i)}}-\beta_{j0}-\beta_{ji}\boldsymbol{x}_{li}-\sum_{\text{m s.t. m is a parent of j \& }i\neq\text{m}}\boldsymbol{\beta}_{jm}\boldsymbol{x}_{lm})^2) \\
&= exp(\frac{1}{2\sigma^2_{i|pa(i)}}(\boldsymbol{x}_{i_{miss}}-\boldsymbol{\beta}_{i|pa(i)}\boldsymbol{x}_{pa(i)_{miss}})^2 \\
&+ \sum_j \frac{1}{2\sigma^2_{j|pa(j)}}(\boldsymbol{x}_{j_{miss(i)}}-\beta_{j0}-\boldsymbol{\beta}_{j|pa(j)(i)}\boldsymbol{x}_{pa(j)(miss_i)}-\sum_m \boldsymbol{\beta}_{m(j)}\boldsymbol{x}_{m(j)})^2)
\end{aligned}
\tag{6.8}
$$

substituting in $\boldsymbol{x}_{pa(j)(miss_i)}$ with $\boldsymbol{x}_{i_{miss}}$ as they are equivalent, and taking expressions only dependent on $\boldsymbol{x}_{i_{miss}}$ I obtain

$$
\begin{aligned}
p(\boldsymbol{x}_{i_{miss}}|\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\sigma}^2,d) \;\propto\; & exp(\boldsymbol{x}^2_{i_{miss}}(\frac{1}{\sigma^2_{i|pa(i)}}+\sum_j \frac{\boldsymbol{\beta}_j|pa(j)(i)}{\sigma^2_{j|pa(j)}}) \\
- \;& 2\boldsymbol{x}_{i_{miss}}(\frac{\boldsymbol{\beta}_i|pa(i)\boldsymbol{x}_{pa(i)_{miss}}}{\sigma^2_{i|pa(i)}}+\sum_j \frac{\boldsymbol{\beta}_{j|pa(j)(i)}}{\sigma^2_{j|pa(j)}}(\boldsymbol{x}_{j_{miss(i)}}-\beta_{j0}-\sum_m \boldsymbol{\beta}_{m(j)}\boldsymbol{x}_{m(j)})))
\end{aligned}
\tag{6.9}
$$

where $\boldsymbol{x}_i$ is the phenotype I am imputing missing values for, $\boldsymbol{x}_{pa(i)_{miss}}$ is the matrix of parents of i corresponding to the missing values of $\boldsymbol{x}_i$, similarly $\boldsymbol{x}_{pa(j)_{miss}}$ represents the column relating to i as a parent of j of the matrix of all parents of j; and specifically the elements of that column associated with the missing elements of $\boldsymbol{x}_i$, $x_{li_{miss}}$ represents each element of missing phenotype $\boldsymbol{x}_{i_{miss}}$ for $l = 1, \ldots, n_{i_{miss}}$, β_{i0} is the β coefficient associated with the intercept value of $\beta_{i|pa(i)}$, $x_{lj_{miss}}$ represents each element of $\boldsymbol{x}_j$ (where j is a child of i) corresponding to missing phenotype $\boldsymbol{x}_{i_{miss}}$ for $l = 1, \ldots, n_{i_{miss}}$, β_{j0} is the β coefficient associated with the intercept value of $\beta_{j|pa(j)}$, $\boldsymbol{\beta}_{jm}$ is the vector of βs associated with parents of j other than i, $x_{lm_{miss}}$ represents each element of $\boldsymbol{x}_m$ (where m is a parent of j not equal to i) corresponding to missing phenotype $\boldsymbol{x}_{i_{miss}}$ for $i = 1, \ldots, n_{i_{miss}}$.

This equates to a Gaussian distribution for $\boldsymbol{x}_{i_{miss}}$, and each missing value of each node i is sampled from:

$$N\left(\frac{\frac{\boldsymbol{\beta}_{i|pa(i)}\boldsymbol{x}_{pa(i)_{miss}}}{\sigma^2_{i|pa(i)}} + \sum_j \frac{\beta_{j|pa(j)(i)}}{\sigma^2_{j|pa(j)}}\left(\boldsymbol{x}_{j_{miss(i)}} - \beta_{j0} - \sum_m \boldsymbol{\beta}_{m(j)}\boldsymbol{x}_{m(j)}\right)}{\frac{1}{\frac{1}{\sigma^2_{i|pa(i)}} + \sum_j \frac{\beta_{j|pa(j)(i)}}{\sigma^2_{j|pa(j)}}}}, \frac{1}{\frac{1}{\sigma^2_{i|pa(i)}} + \sum_j \frac{\beta_{j|pa(j)(i)}}{\sigma^2_{j|pa(j)}}}\right) \tag{6.10}$$

6.3.8 Binary data

The algorithm was then extended to allow for binary phenotype data. Via data augmentation I can model binary regression with a probit link, using essentially a Gaussian linear model. [25, 71].

As with the data augmentation in binary probit regression in the BMARS chapter, consider

$$\boldsymbol{x}_i \sim Bernoulli(\Phi(\eta_i)) \tag{6.11}$$

where $\boldsymbol{x}_i$ is a binary phenotype. Note: $\boldsymbol{x}_i$ can be either a proposed child or parent here.

Introducing a set of latent variables z_i for the i^{th} observation where

$$z_i \sim N(\eta_i, 1) \tag{6.12}$$

such that

$$x_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

In the case of my algorithm, at each iteration, $\boldsymbol{z}$ was sampled from a truncated Gaussian distribution with mean of expected value of each element given the distribution of the current DAG. i.e. the mean value given current values of all parents, children and related parameters of current DAG for each binary node. Taking into account the joint distribution of the current DAG model

$$\eta_i = \frac{\frac{\beta_i x_i}{\sigma_i^2} + \sum_j \frac{\beta_{ji}}{\sigma_j^2}(x_j - \beta_{j_{int}} - \sum_m \beta_{j_m} x_{j_m})}{\frac{1}{\frac{1}{\sigma_i^2} + \sum_j \frac{\beta_{ji}^2}{\sigma_j^2}}} \tag{6.13}$$

using the same principle as for missing phenotype data.

So I sample z_i from a truncated Gaussian distribution defined by

$$z_i \sim N(\eta_i, 1)\,[I > 0] \tag{6.14}$$

6.4 Assessing Performance of the Algorithm

The algorithm was run for several different numbers of nodes, using simulated data at first to check the algorithm was working correctly, and then using data from the NPHS-II study from which I selected variables which already had known biological associations to further test the algorithm.

6.4.1 Simulation study

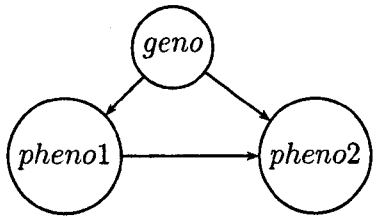
The first step of testing the algorithm, was to use simulated data. I used up to 6 nodes with up to two simulated genotypes, and up to 5 simulated phenotypes for 1,000 individuals. The genotype was simulated by randomly sampling the number 0,1, or 2 1,000 times to correspond to the 3 possible genotypes with probabilities under HWE with MAF ≥ 0.05. For my first simulation with 3 nodes, the phenotypes were simulated by setting

$$\text{phenotype1} = 0.7 + (2*\text{genotype}) + \alpha_1$$

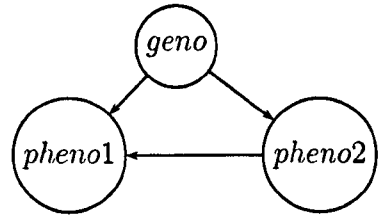
$$\text{phenotype2} = 0.5 + (1.5*\text{genotype}) + (2.5 * \text{phenotype1}) + \alpha_2$$

where α_1 and $\alpha_2 \sim N(0,1)$. Note: These simulated effect sizes are large in comparison to the small genotypic effects expected on phenotypes.

The chains converged quickly within 100 iterations, and settled down between the two DAGs of the same equivalence class within 2000 iterations. The posterior probabilities of the resulting DAGs showed that associations between the nodes were exactly as expected. The results gave the following DAGs and posterior probabilities (after burn-in period):



Posterior probability of 0.41



Posterior probability of 0.37

The two DAGs above are in the same equivalence classes and have approximately equal posterior probabilities of 0.41 and 0.37. The marginal posterior probability of a direct association from phenotype 1 to phenotype 2 is approximately equal to that from phenotype 2 to phenotype 1 (≈ 0.5). In addition, the regression coefficients were very close to the true values simulated. The first DAG has a model of

phenotype1 = 0.70 + (2.01*genotype)

phenotype2 = 0.5 0+ (1.47*genotype) + (2.51 * phenotype1)

6.4.2 Checks for convergence

Checks for convergence were done in R using the coda package [72], for all coefficients of directed associations for each node. Convergence plots, and summaries of the coefficients were produced and shown below.

Auto-correlation plots below show that the algorithm is mixing well and there is no dependence between iterations after a lag of 10.

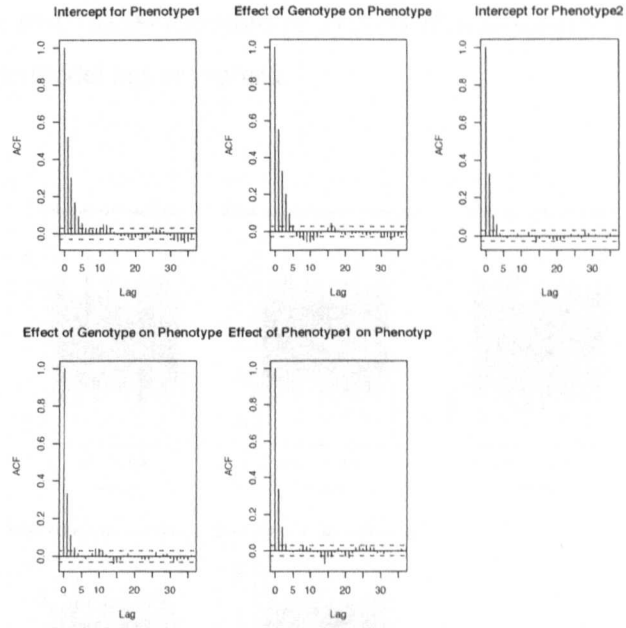


Figure 6.5: Auto-correlation of βs

Density plots of the first 10,000 iterations of each coefficient specific to the DAG simulated show that the model has recovered the coefficients as simulated in each case. (0.7, 2, 0.5, 1.5 and 2 respectively).

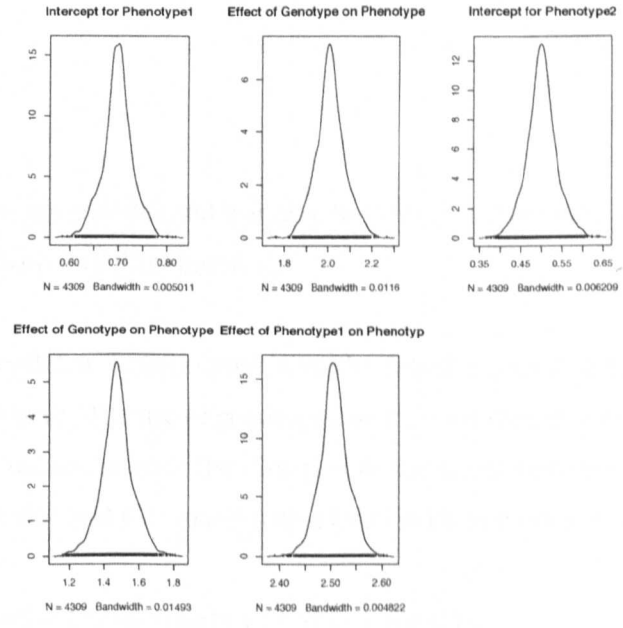


Figure 6.6: Density plots of βs

Trace plots of the first 10,000 iterations of each coefficient specific to the DAG simulated show that the model has converged.

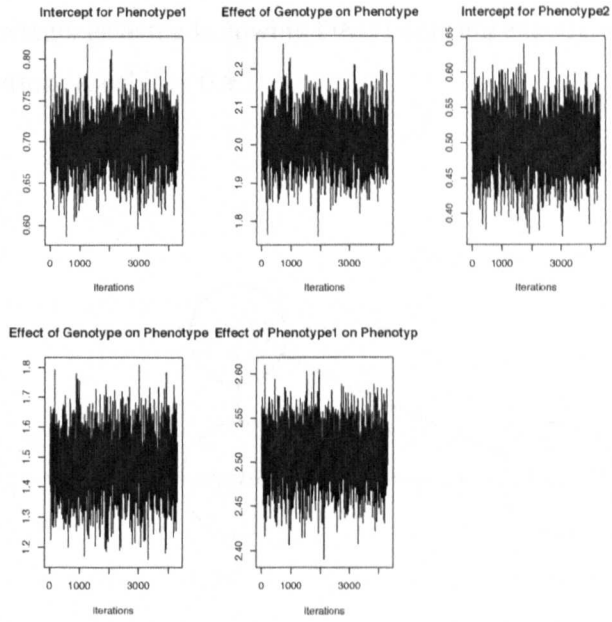


Figure 6.7: Trace plots of βs

I simulated data for more nodes and was able to recover the true structures. I simulated for up to 6 nodes with 2 million iterations.

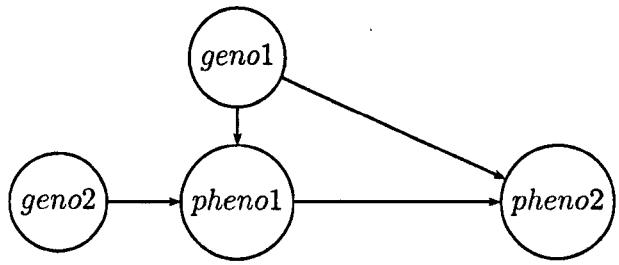
For example, I simulated 3 phenotypes with the following relationships to the genotype simulated as above. The use of genotype 2 as an instrumental variable (see Section 4.8) should allow me to discover the direction of the arrow between phenotype1 and phenotype2 as this will lead to a model with a DAG in its own unique equivalence class.

phenotype1 = 0.7+(2*genotype1) + (1.5*genotype2) + α_1

phenotype2 = 0.5 +(1.5*genotype1)+ (2.5*phenotype1)+ α_2

where α_1 and $\alpha_2 \sim N(0,1)$.

I ran the algorithm for 1,000,000 iterations with a burn-in period of 100,000 and a thin of 100. The results gave the following DAG (showing exactly what was modelled above) with posterior probability 0.92.



The same convergence plots as for the 3 nodes simulation above showed that the algorithm mixed well, the true correlation coefficients were recovered and the algorithm had converged after 50,000 iterations. As there is no DAG equivalent to this one, the direction of association (as that simulated) between phenotype 1 and phenotype 2 is clear due to the addition of genotype 2 as an instrumental variable.

In addition, a table of the marginal probabilities of each arrow shows that the algorithm selects a directed association from genotype 1 to phenotypes 1 and 2 with posterior probability 1, from genotype 2 to phenotype 1 with probability 1, from genotype 2 to phenotype 2 with probability 0.09, and from phenotype 1 to phenotype 2 with probability 1, as expected.

child / parent	phenotype 1	phenotype 2
genotype 1	1	1
genotype 2	1	0.09
phenotype 1	0	1
phenotype 2	0	0

Table 6.1: Marginal probabilities of directed association

6.4.3 Application to real data

Model 1

I then applied my algorithm to real data. I focused on using previously known associations described in Drenos et al. [63]. As summarised in tables above there is a known association between APOB levels with APOB and APOE genotypes. CRP levels are also associated with APOE genotype but also with CRP genotype and APOB level. Also, smoking is highly associated with CRP level, and so it may be interesting to see how these variables interact jointly with the addition of smoking. As with Drenos et al, I have used log CRP and log APOB levels, as these have Gaussian distributions.

How these are related jointly will be interesting. It is expected biologically [73] that the association between APOE genotype and CRP level is only through confounding, and once adjusting jointly for other genotypes, this association will not be apparent. It is unknown what the direction of association between APOB level and CRP level is. Smoking is known only to be associated with levels of CRP, and modelling this relationship jointly with everything else, should not affect the relationship between the other variables. An additive APOE genotype model is associated with increased levels of APOB level but decreased levels of CRP. APOB genotype is known to be associated with increased levels of APOB, but has no affect on CRP levels. In this subset of variables, CRP genotype is only known to be associated with increased levels of CRP.

I ran my algorithm with these 5 variables as nodes. The data set included 2,385 individuals of which 1895 individuals had no missing data at any of the 5 variables. As shown in Section 6.2, there is no association between any of the genotypes, and therefore no

associations between genotypes were allowed in the matrix of allowed direction. The algorithm was run for 1 million iterations with a burn-in period of 100,000 iterations.

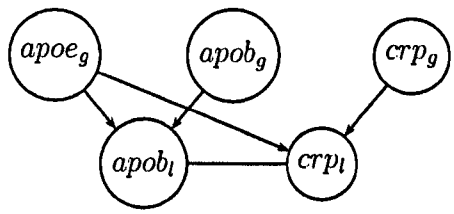
Note: As APOE genotype is a two SNP variant, it was coded -1, 0, 1 for $\varepsilon 2^*$, $\varepsilon 33$ and $\varepsilon 4^*$ respectively.

The table below shows the p-values associated with independent additive models between the 5 above variables and smoking status.

	APOB genotype	CRP genotype	APOE genotype	CRP level	APOB level
CRP level	0.28	4.05E-03	5.25E-05		4.25E-07
APOB level	1.60E-04	0.46	4.72E-10	4.25E-07	
Smoking	0.519	0.488	0.22	<2E-16	0.86

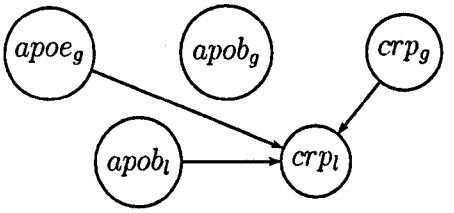
Table 6.2: p-values for independent tests of association

If all the independent associations shown above hold when the 5 variables (not including smoking, initially) are modelled **jointly,** I would expect the following DAG

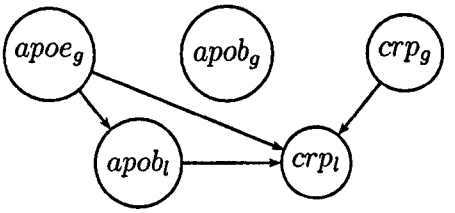


I ran the model 10 times for 1 million iterations with a burn in period of 100,000 after which the model had converged. Plots of convergence are shown in the Appendix 9.1. The acceptance rate of a birth, death or switch step in the MCMC was 9.4%. This means that the algorithm is mixing well.

After a burn-in period of 100,000 iterations, the DAGs with the largest posterior probabilities are as follows:



Posterior probability of 0.18



Posterior probability of 0.15

These DAGs are both unique in terms of equivalence classes. Therefore, they have their own underlying joint distributions.

Note: the next top models have posterior probabilities of 0.11, 0.10, 0.07 and 0.04.

The only difference between the two top models, is the addition of an association between between APOE genotype and APOB level. Considering these joint posterior probabilities alone, this perhaps gives some indication that the relationship between APOE genotype and CRP level is through APOB level, once modelled jointly with the other variables. However, the marginal posterior probabilities appear to give more information about the underlying associations.

The marginal probabilities of directed association are shown in the table below. There are marginal posterior probabilities of 0.999 and 1 that CRP genotype and APOE genotype are associated with CRP level respectively. There is a marginal posterior probability of 0.45 that APOB level is dependent on APOE genotype. There is a marginal posterior probability of 0.49 that CRP level is dependent on APOB level; and a marginal

posterior probability of 0.33 that APOB level is dependent on CRP level.

child parent	APOB level	CRP level
CRP genotype	0.06	0.999
APOE genotype	0.45	1
APOB genotype	0.14	0.17
APOB level	0	0.49
CRP level	0.33	0

The results in this table suggest that i) CRP genotype only effects CRP level, ii) CRP level is definitely dependent on APOE genotype and iii) there is a relationship between APOB level and CRP level. There is weaker evidence that APOB level is dependent on APOE genotype. The direction of the association between APOB level and CRP level is more likely to be from APOB level to CRP level. Interestingly, in this model, APOB genotype appears to have little (or no) effect on any other variable, including APOB level.

The model with the highest posterior probability (0.18) is log CRP level = 1.1 + (0.26* CRP genotype) - (0.19* APOE genotype) + (0.34 * log APOB level).

In the joint model with all five variables, after adjusting for APOE genotype and CRP genotype on CRP level there is no association between APOB genotype with APOB level. The relationship of APOB level on the other variables in the model may be through CRP level but perhaps not considering . In order to determine the direction of association between APOB level and CRP level, other nodes could be added that may act as instrumental variables, or explain in more detail the relationship.

It is strange that the marginal association between APOB genotype and APOB level is so low. This could be because the relationship between APOE genotype, CRP level and APOB level is so strong that once adjusted for jointly, the relationship is no longer significant. There is something strange underlying this model but the biology behind it is unknown. There is not enough information in this model to explain the true rela-

tionships.

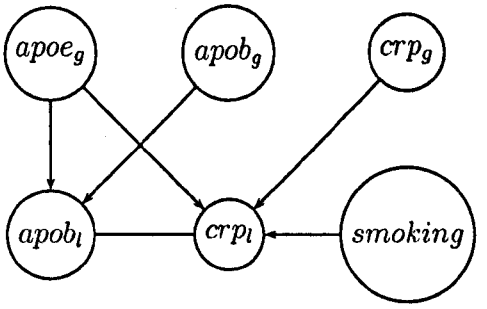
One potential idea is that APOE genotype and APOB genotype are related in some way as this would complicate the relationships shown. However, they are not in LD or on the same chromosome. Biologically, it is though that the association between APOE genotype and CRP level is due to confounding. Once adjusting for other genotypes in LD with APOE genotype, it is thought that this relationship would not exist. However, I do not have these genotypes available, and my model can not cope with larger numbers of variables easily.

In the absence of other information, another possibility is that the associations between APOE genotype and CRP level are not, in fact, modulated through APOE level, for example. I do not, however, have APOE level data and conditioning on this variable could possibly be sufficient to identify the direction of association between APOB level and CRP level. Another possibility to identify the direction of association between APOB level and CRP level in this joint model is by using a confounder between the two. In my dataset, smoking is associated with CRP level but not APOB level, and so cannot be used as a possible confounder in the joint model. However, it is still interesting to investigate how adding smoking to the model will effect the joint model.

6.4.4 Model 2

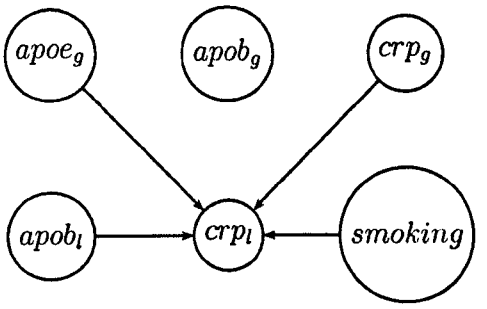
I ran the above model but with the addition of a smoking variable. I restricted the arrow between smoking and phenotype APOB level or CRP level to be directed from smoking to phenotype. It is unlikely that phenotype level will have a direct effect on smoking status. Smoking level is strongly associated with CRP level with a p-value of <2E-16. It is also unlikely that genotype will have a direct effect on smoking status. As shown in Table 6.2, there is no association between smoking status and any of the genotypes in this model. (p-values of 0.52, 0.49 and 0.22 for an association between smoking status and APOB genotype, CRP genotype and APOE genotype respectively). Therefore, I did not allow for an arrow directly between genotype and smoking. I ran my algorithm 10 times for 1 million iterations, a thin of 100 and a burn in of 100,000.

If all the independent associations shown above in Table 6.2 hold when the 6 variables are modelled **jointly,** I would expect the following DAG

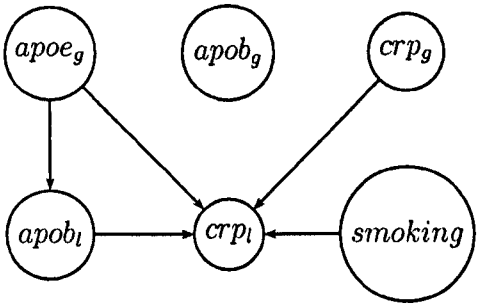


Plots of convergence are shown in Appendix 9.2. The acceptance rate of a birth, death or switch step being accepted in the MCMC was 8.4%

The DAGs with the highest posterior probabilities were obtained for the following joint models



Posterior probability of 0.16



Posterior probability of 0.14

The joint relationships have similar posterior probabilities as to the model without smoking but with the added relationship between smoking and CRP level.

Again, these DAGs have different underlying joint distributions and are members of unique equivalence classes.

Considering the marginal posterior probabilities in the table below, it is evident that the addition of smoking to the model does not change much.

This is reassuring in terms of the prior on model size as the increase in variables has not altered the results significantly.

child / parent	APOB level	CRP level
CRP genotype	0.06	0.98
APOE genotype	0.48	1
APOB genotype	0.13	0.15
Smoking	0.08	1
APOB level	0	0.49
CRP level	0.34	0

The joint model with the highest posterior probability (0.16) is log CRP level = 1.0 + (0.23* CRP genotype) - (0.18* APOE genotype) + (0.31 * APOB level) + (0.5 *Smoking status).

6.4.5 Model 3

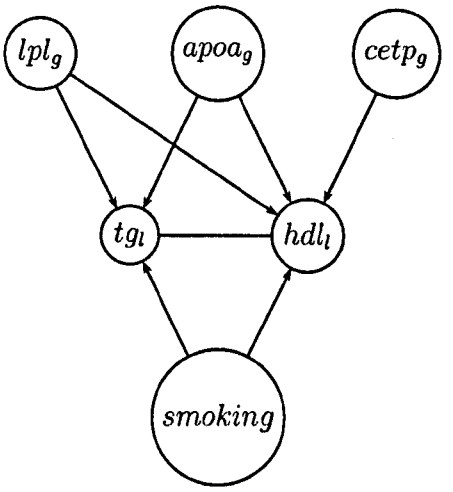
Another set of variables with potentially interesting joint relationships as highlighted by Drenos et al [63] are CETP genotype, APOA genotype, LPL genotype, HDL level, TG level and smoking status. Smoking status is a known confounder of the relationship between HDL level and TG level. LPL and APOA genotypes are associated with both TG and HDL levels. CETP genotype is associated with HDL level. I used my algorithm to investigate how these relationship work jointly.

331 individuals out of the dataset of 2,385 were missing HDL level, and 13 were missing TG level.

The p-values associated with independent relationships between these variables is shown in the table below.

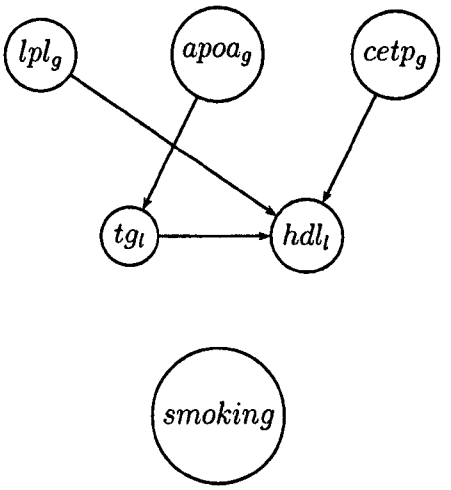
	CETP genotype	APOA genotype	LPL genotype	HDL level	TG level
HDL level	0.00117	0.0114	0.000228		<2.16E-16
TG level	0.283	8.09E-05	0.00263	<2.16E-16	
Smoking	0.50	0.94	0.97	0.0306	0.045

If all these independent associations occurred jointly, I would expect the following DAG assuming a 5% significance level.

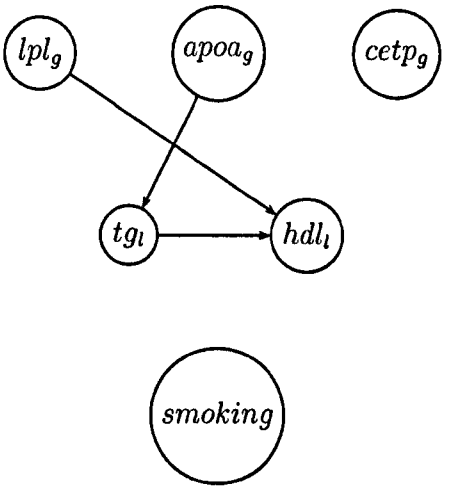


I ran the model 10 times for 1 million iterations with a burn-in of 100,000 iterations each. The acceptance rate of a birth, death or switch step being accepted in the MCMC was 8.3% Plots of convergence are shown in Appendix 9.3.

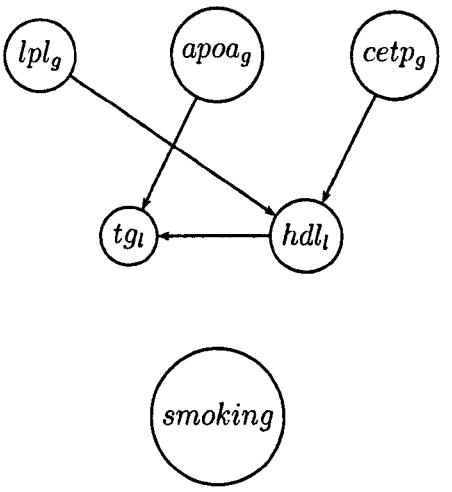
The DAGs with the highest posterior probabilities were obtained for the following joint models



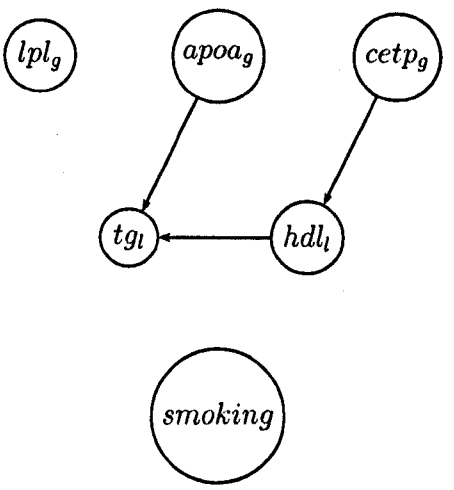
Posterior probability of 0.11



Posterior probability of 0.09



Posterior probability of 0.09



Posterior probability of 0.08

Note: Joint top models shown are those that are most common. In this case no clear model seems to have come up.

Marginal posterior probabilities of direct association between the 6 variables is shown in the table below.

child / parent	HDL level	TG level
LPL genotype	0.53	0.12
CETP genotype	0.83	0.04
APOA genotype	0.15	0.88
HDL level	0	0.49
TG level	0.51	0
Smoking	0.20	0.12

In this case, my algorithm does not seem to select a particular model with high posterior probability. The DAGs with the highest posterior probabilities indicate that the association between APOA genotype and HDL level is modulated through TG level; and the relationship between LPL genotype and TG level is modulated through HDL level. Again, the marginal posterior probabilities give more information and in this case re-iterate what is seen in the joint models. There is a posterior probability of 0.53 of a direct association between LPL genotype and HDL level compared to 0.12 of a direct association between LPL genotype and TG level. Similarly, there is a posterior probability of 0.88 of a direct association between APOA genotype and TG level whereas there is a posterior probability of 0.15 of a direct association between APOA genotype and HDL level.

The marginal probabilities found suggest that the relationship between CETP genotype and TG level is modulated through HDL level. It appears that

i) HDL level is dependent on CETP genotype
ii) HDL level is dependent on LPL genotype
iii) TG level is dependent on APOA genotype

iv) There is a relationship between TG level and HDL level

Biologically, it may have been expected that the algorithm would not pick a particular model as the relationship between TG level and HDL level are so highly correlated.

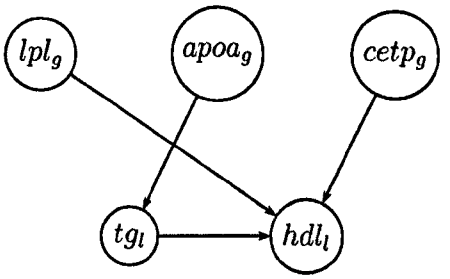
The joint model with the highest posterior probability (0.11) is HDL level = 1.94 - (0.07 * LPL genotype) + (0.09 * CETP genotype) - (0.45 * log TG level) where log TG level = 0.56 + (0.11* APOA genotype)

The DAGs with the highest posterior probabilities shown above all have unique equivalence classes and therefore different underlying joint distributions. However, they all have similar posterior probabilities so it is difficult to distinguish from them what the true joint relationships are. The marginal posterior probabilities give more information and given these, the only important direction that is not defined is that between HDL level and TG level. It is likely that there are more unknown factors involved in this highly correlated relationship. Again, my model is not able to include any more variables easily even if they were available.

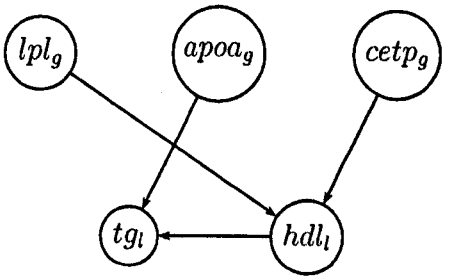
As there is weak evidence of a joint model including an association between smoking and either HDL or TG level, I dropped smoking from the dataset.

6.4.6 Model 4

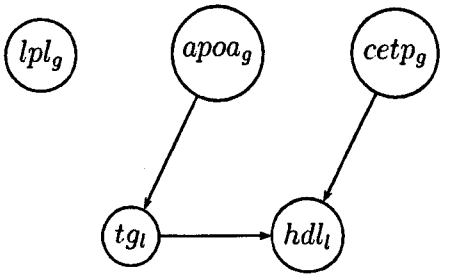
I ran my algorithm on the data used in Model 3 but without smoking status. Again, I ran the algorithm 10 times for 1 million iterations and a burn-in period of 100,000. Acceptance rate was 9%. Convergence plots are shown in Appendix 9.4. The DAGs with the highest posterior probabilities are shown below.



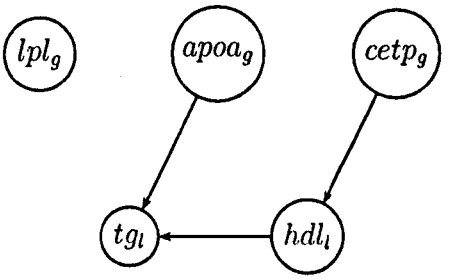
Posterior probability of 0.15



Posterior probability of 0.13



Posterior probability of 0.13



Posterior probability of 0.13

Marginal posterior probabilities of direct association between the 5 variables is shown in the table below.

parent \ child	HDL level	TG level
LPL genotype	0.53	0.13
CETP genotype	0.83	0.04
APOA genotype	0.15	0.88
HDL level	0	0.49
TG level	0.51	0

Without smoking in the model, there is a smaller subset of possible arrows in the DAG, and therefore the posterior probabilities of a particular model are not spread so thinly. As seen here, the posterior probabilities of the same DAGs increase by 0.02 compared to those in Model 3. The marginal probabilities remain the same as expected.

6.5 Conclusions and Discussion

6.5.1 Simulated Data

As demonstrated by analysis on simulated datasets, my algorithm works well on small numbers of nodes and arrow subsets. Approximately equal posterior probabilities are obtained for simulated equivalence classes as shown by the simulated data example with 3 nodes. Each equivalence class has a posterior probability of ≈ 0.4 and the marginal posterior probability of phenotype 1 being dependent on phenotype 2 is equal to phenotype 2 being dependent on phenotype 1 (≈ 0.5).

The addition of an instrumental variable in the form of genotype 2 to the simulated dataset gives the correct DAG with a high posterior probability of 0.92. From this DAG is is possible to recover the direction of association between phenotype 1 and phenotype 2 which is recovered as modelled i.e. phenotype 2 is dependent on phenotype 1. This DAG is in its own equivalence class and so this has helped me to determine the correct/ simulated direction of association as discussed in introduction Section 4.8.

My algorithm also recovered the coefficients as simulated. The convergence plots show that the algorithm is mixing well and converge to the "correct" model quickly.

6.5.2 NPHS-II Data

The results on analysis using the NPHS-II data are, as expected, not so clear as those using the simulated data.

Model 1 was an analysis on the joint relationships between APOE genotype (rs429358 + rs7412), APOB genotype (rs585967), CRP genotype (rs3091244), APOB level and CRP level. The results gave two DAGS with high posterior probabilities of 0.18 and 0.15. These showed that when examining the relationships between all these variables jointly, only 3 direct effects are important and 1 of possible importance once adjusting for the effects of all other variables in the model. The relationship between APOB genotype and APOB level; and APOE genotype and APOB level were not important in the joint model. These results suggest surprisingly that the relationship between APOE genotype and APOB level appears to be modulated through CRP level. As the two DAGs shown above for Model 1 are unique in terms of equivalence classes, I can conclude that there is a slightly higher posterior probability of CRP level being dependent on APOB level rather than the other way round, when taking into account the underlying joint distribution with the other variables.

Looking at the marginal posterior probabilities, these probabilities re-iterate this: there is a marginal posterior probability of 0.49 of a direct association of APOB level on CRP level but only 0.33 of a direct effect of CRP level on APOB level. It is interesting that the independent p-value of association between APOB level and APOB genotype of 1.60E-04 compares to 0.14 marginal posterior probability when the relationship is examined in a joint model. i.e. APOB level regressed upon both APOB genotype and CRP level with all other variables controlled for. Also, APOE genotype has an independent frequentist p-value of association with APOB level of 4.72E-10. This relationship has a marginal posterior probability of 0.45 under the DAG framework.

Model 2 used the same variables as Model 1 but with the addition of smoking status. Given the strong frequentist association found between CRP level and smoking

independently, it is not surprising that there is a marginal posterior probability of 1 between these. Other than the addition of a direct association of smoking on CRP level, the results from Model 2 change only a tiny amount from those found in Model 1. The relationship between smoking and CRP level does not affect the joint relationships between APOE genotype, CRP genotype, APOB level and CRP level.

I examined the joint relationships between LPL genotype (rs264), APOA genotype (rs5689566), CETP genotype (rs708272), HDL level, TG level and smoking status in Models 3 and 4. Smoking status is a known confounder between HDL level and TG level so it was interesting to see that smoking has no relationship with either in this model after jointly adjusting for the genotypes (in the DAGs with the highest posterior probabilities reported above). Using my DAG methodology, smoking status had a marginal posterior probability of 0.2 and 0.12 of direct association with HDL level and TG level respectively. These relationships had independent p-values of 0.03 and 0.045.

The results from this model show that the effect of APOA genotype on HDL level is modulated through TG. The effect of LPL genotype on TG level is modulated through HDL genotype. Again, considering the marginal posterior probabilities, the relationship between CETP genotype and TG level is also modulated through HDL level. The independent frequentist tests of association between APOA genotype and HDL level and TG level have p-values of 0.01 and 8.09E-05 receptively compared to marginal posterior probabilities of 0.15 and 0.88 when modelled jointly. Similarly, the p-values of independent association between LPL genotype and HDL and TG levels are 0.0002 and 0.003 compared to marginal posterior probabilities of 0.53 and 0.12.

The direction of association between HDL level and TG level is still undetermined. It may be possible to establish the direction of this association if an appropriate variable was discovered that could act as an instrumental variable. As described in Section 4.8 a variable is instrumental in this case if it is only associated with one variable through the other. However, I do not know of any such variables at present.

Alternatively, as HDL level and TG level are so highly correlated, better plasma traits could be used. HDL level, for example, is composed of many lipids and these could

be incorporated.

6.5.3 Discussion

The methodology used in my joint model analysis has automatically corrected for the effects of all variables in the model via the edges defined. By allowing the algorithm to choose the appropriate model any confounders in the data are automatically corrected for as well.

Given these results, it is still tricky to determine exactly what the directed joint relationships are between the variables used is. Genotype information has helped to establish some relationships. My Bayesian methodology gives a better idea of what the underlying joint distributions are by quantifying the posterior probability of each DAG through model selection. This adds more information than simple independent tests of association or joint frequentist models. [74]

The above results are in the absence of any other information. It would, for example, be interesting to examine the effects of APOE level (which is currently unavailable in the NPHS-II dataset) on the joint model between APOE genotype, APOB genotype, CRP genotype, CRP level and APOB level. Incorporating APOE level into this dataset may alter the relationship between the other variables jointly. It could be, if APOE level is strongly correlated with CRP and APOB levels, that once adjusting for APOE level there is no direct association between APOE genotype and CRP level. This effect could be modulated through APOE level.

In addition, biologically it is thought the relationship between APOE genotype, APOB level and CRP level is more complex than a model with only 3 genotypes. The inclusion of other genotypes in LD with APOE genotype may explain the relationship more adequately and answer the question as to whether the apparent relationship between APOE genotype and CRP level is only due to confounding.

It is also important to note that it is strange that in this model, the relationship between

APOB genotype and APOB level is not important. There is something underlying this but it is not known what. I do not know enough about the biology behind it, or have enough information in terms of other variables to establish anything other than the results shown. I believe that the methodology used works but there is not enough in the model/ data I was given to establish sensible results. The relationship between APOE genotype, APOB genotype, CRP genotype, APOB level and CRP level is more complex than first expected.

My algorithm is limited by the number of nodes and subset of possible arrows in the DAG model. By expanding/ developing my code in another program (C++, for example), I could achieve more computing power. The computing time for my analysis is approximately 24 hours for 6 nodes for 1million iterations. Again, this could be improved by using C++. If I were able to increase efficiency greatly then I would be able to include more possible confounders/ variables considered potentially biologically important in the model.

In order to further investigate the efficacy of my methodology, simulated data with effect sizes more similar to the typically small genetic effects expected could be used. This would help determine whether the inconclusive results are due to the data or my algorithm not having enough power to detect small effect sizes.

In order to better understand these relationships, I think the next development would be to incorporate prior biological knowledge on pathway analysis. This may help to disentangle true relationships from spurious ones. Gene pathways are now well characterised with relevant information publicly available on databases such as KEGG [75] and Gene Ontology [76].

CHAPTER

7

DISCUSSION

The aim of this PhD was to try to understand the factors influencing disease by trying to disentangle genetic and phenotypic information. I have developed a new method for jointly modelling genotypes and phenotypes in an attempt to disentangle complicated relationships; and applied an already existing method to find causal SNPs for SLE from those in very high LD. Both of these were looking for a more simple structure underlying the data and were achieved using a Bayesian framework.

7.1 Bayesian Multivariate Adaptive Regression Spline Modelling

I applied the Bayesian Multivariate Adaptive Regression Spline model developed by Verzilli et al. [59] to find causal SNPs for SLE amongst those in high LD on the MHC region. This expanded on work by Rioux et al [46] who identified primary association signals, and then performed conditional regression to identify independent sec-

ondary signals. They reported at least three separate signals using this method, namely RS1269852, RS3135391 and the NOTCH gene. My method highlighted at least four independent signals with posterior probability 0.72. These were either RS558702 or RS1296852 with either RS3135391 or RS135388 in addition to two other signals. There was strong evidence for two other signals along the chromosome but without high posterior probabilities of specific SNPs. Therefore, the BMARS model allowed me to identify further SNPs and regions of interest for further investigation. In addition, the BMARS methodology was more flexible as it allowed for genotype interactions and different effects for each SNP.

7.1.1 Comparison to Other Methods

Other current methods of determining SNP association in this respect include frequentist stepwise regression or the Bayesian method, Bayesian IMputation-Based Association Mapping (Bim-Bam) [77].

Frequentist stepwise regression would only result in one model with no known certainty that it is the correct model. The Bayesian methodology used gives quantitative posterior probabilities of each model; of the number of signals; and of each SNP having an association. It is unlikely that a single model will represent the data. It is more likely that a number of similar models will be almost equally representative of the underlying model. BMARS quantifies the posterior probability of each of these models using model averaging whereas a frequentist stepwise approach forces the choice of a single model.

The Bim-Bam method shares a few similarities with BMARS. Bim-Bam is also a method to detect causal SNPs when multiple causal variants are present. Bim-Bam uses LD information together with typed genotypes to impute missing genotypes, and then uses Bayesian regression to assess association between the phenotype and the estimated genotypes. Bim-Bam then quantifies the strength of evidence that each SNP is associated with Bayes Factors. Therefore, like the method used in this PhD, this approach provides more interpretable explanations for observed associations, compared

to frequentist methods or single SNP tests.

However, BMARS has several advantages over Bim-Bam in that it allows for interactions between genotypes and uses a reversible jump algorithm. The use of interactions more accurately defines the relationships between SNPs and the outcome. The use of reversible jump means that a BMARS approach results in model selection and actually eliminates SNPs from the model whereas Bim-Bam summarises the strength of association of all SNPs individually. This means that with BAMRS the strength of evidence for each model is quantified with a posterior probability; rather just the strength of association for each single SNP. By selecting the most likely models, BMARS is modelling SNPs jointly and so selects the true causal SNPs from those in high LD.

7.2 Bayesian Networks for Genetic Association Studies

The methodology I developed on Bayesian networks for genetic association studies modelled several variables jointly; automatically correcting for the effects of all variables in the model via the edges defined. This work expanded on the independent tests of association between SNPs and phenotypes highlighted by Drenos et al. [63] in the NPHS-II data. The analysis I carried out gave more information about these associations by allowing for the joint underlying distributions. Again, as I have used a Bayesian framework, posterior probabilities of each model and marginal posterior probabilities of association between each set of variables are achieved. The DAG framework also allows the possibility to determine the direction of associations between variables.

7.3 Advantages of Methodology Used

Both methods in my PhD thesis make use of Bayesian statistics. In a context where there is a lot of uncertainty about which SNP is the most causal out of those in LD, or which association is the strongest given weak genetic effects, this has several advantages as discussed above. i.e. strength of evidence quantified in posterior probabilities

of models and marginal probabilities of associations. The priors used in my models were uninformative and sensitivity analysis showed that the priors did not affect the results.

7.4 Future Work

Both methods are attempting to explain complex genetic relationships. I believe that the future of genetic research is in gene-environment interplay. More information and methodology that allow for highly correlated factors in analysis are needed. The key to how genetics affects diseases is in understanding how genes and environmental factors are interlinking. Fine mapping projects using re-sequencing are emerging such as the 1,000 Genomes project [23]. This is an example of biology and genome-wide association studies brought together. These projects collect data with complicated pathways and there are likely to be a lot of SNPs in high LD interacting with other biological factors to be disentangled. It is becoming more and more apparent that genes act with other biological factors in complex pathways.

The BMARS methodology used in this PhD, in my opinion, could be helpful in identifying causal SNPs (or clusters of SNPs) for further analysis in these large, complex datasets. The BMARS model already allows for epistatis, in that it is very flexible in allowing for interactions between genotypes and different genotypic effects on the outcome (due to the spline feature for dominance, additivity, etc.). The BMARS methodology could also easily be extended to include other factors such as multiple phenotypes. I believe that BMARS is a computationally efficient way to establish causal SNPs from large datasets.

My methodology for networks could be expanded to allow for more variables computationally, to investigate further these complex relationships. With more and more data being collected, we are likely to have larger sample sizes, which would be better for detecting small genotypic effects due to an increase in power. This would be useful to disentangle the many questions we have about gene-environment interaction, and direction of effects. The variables would be modelled jointly and hopefully find true

effects from spurious ones once everything in the model has been adjusted for. The methodology could also be extended to include nodes for genotype-phenotype interaction.

BIBLIOGRAPHY

[1] Alberts, Johnson, Lewis, Raff, Roberts, and Walter. *Molecular Biology of The Cell.* Garland Science, New York, 4th edition, 2002.

[2] Griffiths, Miller, Suzuki, Lewontin, and Gelbart. *An Introduction to Genetic Analysis.* W.H. Freeman and Company, New York, 7th edition, 2000.

[3] S.M. Brown. *Essentials of Medical Genomics.* Wiley Online Library, 2003.

[4] Date accessed 25th june 2011. www.biotechnologyonline.gov.au.

[5] Date accessed 25th june 2011. http://www.accessexcellence.org/RC/VL/GG/images/comeiosis.gif.

[6] D.J. Balding, M. Bishop, and C. Cannings. *Handbook of Statistical Genetics.* Wiley, Chichester, England, 3 edition, 2007.

[7] Date accessed 25th june 2011. http://www.nature.com/nrg/journal/v3/n4/images/nrg777-f1.jpg.

[8] The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–96, 2003.

[9] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 2007.

[10] The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, 2010.

[11] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, and et.al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, 2001.

[12] K. Zhang, M. H. Deng, T. Chen, M. S. Waterman, and F. Z. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7335–7339, 2002.

[13] X. Ke and L. R. Cardon. Efficient selective screening of haplotype tag snps. *Bioinformatics*, 19(2):287–8, 2003.

[14] P. Sebastiani, R. Lazarus, S. T. Weiss, L. M. Kunkel, I. S. Kohane, and M. F. Ramoni. Minimal haplotype tagging. *Proc Natl Acad Sci U S A*, 100(17):9900–5, 2003.

[15] D. J. Balding. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–91, 2006.

[16] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5), 2009.

[17] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–34, 2010.

[18] Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annu Rev Genomics Hum Genet*, 10:387–406, 2009.

[19] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 2010.

[20] S. R. Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet*, 124(5):439–50, 2008.

[21] Y. F. Pei, L. Zhang, J. Li, and H. W. Deng. Analyses and comparison of imputation-based association methods. *PLoS One*, 5(5), 2010.

[22] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6), 2009.

[23] 1000 genome project. http://www.1000genomes.org/.

[24] R.A. Fisher. On the interpretation of χ^2 from contingency tables, and calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[25] J. H. Albert and S. Chibb. Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[26] A. Gelman, J.B. Carlin, H.S. Stern, and Rubin D.B. *Bayesian Data Analysis*. Chapman and Hall CRC Press, London, 2nd edition, 2004.

[27] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC Press, London, 2nd edition, 1996.

[28] D Gammerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, London, 1st edition, 1997.

[29] J.A. Rice. *Mathematical Statistics and Data Analysis*. International Thompson Publishing, 2nd edition, 1995.

[30] A.J. Dobson. *An Introduction to Statistical Modelling*. Chapman & Hall, London, 1983.

[31] S.R. Eliason. *Maximum Likelihood Estimation : Logic and Practice*. Oxford University Press, Oxford, 1993.

[32] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Liklehood*. Oxford University Press, Oxford, 2001.

[33] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2004.

[34] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[35] R. Weiss. An approach to bayesian sensitivity analysis. *Journal of the Royal Statistical Society Series B-Methodological*, 58(4):739–750, 1996.

[36] J. Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genet Epidemiol*, 33(1):79–86, 2009.

[37] R.E. Kass and A.E. Rafferty. Bayes factors and model uncertainty. *Journal of the American Statistical Society*, 90(430):773–795, 1995.

[38] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B-Methodological*, 59(4):731–758, 1997.

[39] S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiological research. *Epidemiology*, 10(1):37–48, 1999.

[40] E. Castillo, J.M. Gutierrez, and A.S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer, 1997.

[41] D. Nitsch, M. Molokhia, L. Smeeth, B. DeStavola, J.C. Whittaker, and D.A. Leon. Limits to causal inference based on mendelian randomization: A comparison with randomized controlled trials. *Am J Epidemiology*, 163(5):397–403, 2005.

[42] S. Burgess and S.G. Thompson. Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in Medicine*, (30):1312–1323, 2011.

[43] S. Beck, Et.al., and MHC Sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401(6756):921–923, 1999.

[44] W. Klitz, J. C. Stephens, M. Grote, and M. Carrington. Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the hla class ii region. *Am J Hum Genet*, 57(6):1436–44, 1995.

[45] A. Stenzel, T. Lu, W. A. Koch, J. Hampe, S. M. Guenther, F. M. De La Vega, M. Krawczak, and S. Schreiber. Patterns of linkage disequilibrium in the mhc region on human chromosome 6p. *Hum Genet*, 114(4):377–85, 2004.

[46] J. D. Rioux, P. Goyette, T. J. Vyse, L. Hammarstrom, M. M. Fernando, and Et.al. Mapping of multiple susceptibility variants within the mhc region for 7 immune-mediated diseases. *Proc Natl Acad Sci U S A*, 106(44):18680–5, 2009.

[47] Date accessed 25th june 2011. http://www.b58cgene.sgul.ac.uk/.

[48] Date accessed 25th june 2011. http://www.amdec.org/initiatives/team-science/new-york-cancer-project.

[49] S. Wright. Systems of mating. *Genetics*, 6(2):111–178, 1921.

[50] S. Wright. Coefficients of inbreeding and relationship. *American Naturalist*, 56:330–338, 1922.

[51] L. V. Clark and M. Jasieniuk. Polysat: an r package for polyploid microsatellite analysis. *Mol Ecol Resour*, 11(3):562–6, 2011.

[52] K. E. Holsinger and B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting f(st). *Nat Rev Genet*, 10(9):639–50, 2009.

[53] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[54] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–5, 2005.

[55] J. C. Barrett. Haploview: Visualization and analysis of snp genotype data. *Cold Spring Harb Protoc*, 2009(10):pdb ip71, 2009.

[56] P. I. W. de Bakker, M. A. R. Ferreira, X. M. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, 17:R122–R128, 2008.

[57] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

[58] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2001.

[59] C. J. Verzilli, J. C. Whittaker, N. Stallard, and D. Chasman. A hierarchical bayesian model for predicting the functional consequences of amino-acid polymorphisms. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 54:191–206, 2005.

[60] L. Wasserman. Bayesian model selection and model averaging. *J Math Psychol*, 44(1):92–107, 2000.

[61] D.G.T. Denison, C.C. Holmes, B.K. Mallick, and A.F.M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley & Sons, Chichester, 2002.

[62] P.N. Hopkins and R.R. Williams. A survey of 246 suggested coronary risk factors. *Atherosclerosis*, 40:1–52, 1981.

[63] F. Drenos, P. J. Talmud, J. P. Casas, L. Smeeth, J. Palmen, S. E. Humphries, and A. D. Hingorani. Integrated associations of genotypes with multiple blood biomarkers linked to coronary heart disease risk. *Human Molecular Genetics*, 18(12):2305–16, 2009.

[64] C. Wallace, S. J. Newhouse, P. Braund, F. Zhang, M. Tobin, M. Falchi, K. Ahmadi, R. J. Dobson, A. C. Marcano, C. Hajat, P. Burton, P. Deloukas, M. Brown, J. M. Connell, A. Dominiczak, G. M. Lathrop, J. Webster, M. Farrall, T. Spector, N. J. Samani, M. J. Caulfield, and P. B. Munroe. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet*, 82(1):139–49, 2008.

[65] M. S. Sandhu, D. M. Waterworth, S. L. Debenham, E. Wheeler, K. Papadakis, J. H. Zhao, K. Song, X. Yuan, T. Johnson, S. Ashford, M. Inouye, R. Luben, M. Sims, D. Hadley, W. McArdle, P. Barter, Y. A. Kesaniemi, R. W. Mahley, R. McPherson, S. M. Grundy, S. A. Bingham, K. T. Khaw, R. J. Loos, G. Waeber,

I. Barroso, D. P. Strachan, P. Deloukas, P. Vollenweider, N. J. Wareham, and V. Mooser. Ldl-cholesterol concentrations: a genome-wide association study. *Lancet*, 371(9611):483–91, 2008.

[66] D. Melzer, J. R. Perry, D. Hernandez, A. M. Corsi, K. Stevens, I. Rafferty, F. Lauretani, A. Murray, J. R. Gibbs, G. Paolisso, S. Rafiq, J. Simon-Sanchez, H. Lango, S. Scholz, M. N. Weedon, S. Arepalli, N. Rice, N. Washecka, A. Hurst, A. Britton, W. Henley, J. van de Leemput, R. Li, A. B. Newman, G. Tranah, T. Harris, V. Panicker, C. Dayan, A. Bennett, M. I. McCarthy, A. Ruokonen, M. R. Jarvelin, J. Guralnik, S. Bandinelli, T. M. Frayling, A. Singleton, and L. Ferrucci. A genome-wide association study identifies protein quantitative trait loci (pqtls). *PLoS Genet*, 4(5), 2008.

[67] G. Ken-Dror, P. J. Talmud, S. E. Humphries, and F. Drenos. Apoe/c1/c4/c2 gene cluster genotypes, haplotypes and lipid levels in prospective coronary heart disease risk among uk healthy men. *Mol Med*, 16(9-10):389–99, 2010.

[68] E.M. Fronk and P. Giudici. Markov chain monte carlo model selection for dag models. *Sonderforschungsbereich*, 386(221), 2000.

[69] D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterisation of several probability distributions. *Uncertainty in Artificial Intelligence, Proceedings*, pages 216–225, 1999.

[70] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–590, 1976.

[71] E.M. Fronk. Model selection for dags via rjmcmc for the discrete and mixed case. *Sonderforschungsbereich*, 386(271), 2002.

[72] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[73] Personal conversations with juan pablo-casas.

[74] P. Tompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55(4):525–534, 1995.

[75] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[76] Gene ontology website. http://www.geneontology.org.

[77] B. Servin and M. Stephens. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLos Genet*, 3(7), 2007.

[78] T Strachan and A.P. Read. *Human Molecular Genetics*. Oxford:BIOS, 3rd edition, 2004.

[79] A. Ziegler and I.R. Konig. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley, 2nd edition, 2010.

[80] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[81] S. Morgan and C. Winship. *Counterfactuals and Causal Inference- Methods and Principles for Social Research*. Cambridge University Press, 2007.

[82] Clayton short course notes. http://www-gene.cimr.cam.ac.uk./clayton/courses.html.

[83] A.P. Dawid. Fundamentals of statistical causality, 2007. RSS/EPSRC Graduate Training Programme, University of Sheffield.

[84] A. Agresti and B. Hitchcock. Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3):297330, 2005.

[85] J. P. Casas, L. E. Bautista, L. Smeeth, P. Sharma, and A. D. Hingorani. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. *Lancet*, 365(9455):224–32, 2005.

[86] V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330, 2007.

[87] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, 1995.

[88] S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

[89] D. J. Lunn, J. C. Whittaker, and N. Best. A bayesian toolkit for genetic association studies. *Genet Epidemiol*, 58, 2007.

[90] D.P. MacKinnon, A.J. Fairchild, and M.S. Fritz. Introduction to statistical mediation analysis. *Annual Revue of Pshycology*, 11(7):499–511, 2010.

[91] J. Pearl. Coment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[92] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[93] D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[94] G. D. Smith, N. Timpson, and S. Ebrahim. Strengthening causal inference in cardiovascular epidemiology through mendelian randomization. *Annals of Medicine*, 40(7):524–541, 2008.

[95] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, 2001.

[96] J. A. Cooper, G. J. Miller, K. A. Bauer, J. H. Morrissey, T. W. Meade, D. J. Howarth, S. Barzegar, J. P. Mitchell, and R. D. Rosenberg. Comparison of novel hemostatic factors and conventional risk factors for prediction of coronary heart disease. *Circulation*, 102(23):2816–22, 2000.

[97] C. J. Willer, S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, J. Strait, W. L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A. J. Swift, M. A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W. M. Chen, Y. Li, L. J. Scott, P. A. Scheet, J. Sundvall, R. M. Watanabe, R. Nagaraja, S. Ebrahim, D. A. Lawlor, Y. Ben-Shlomo, G. Davey-Smith, A. R. Shuldiner, R. Collins, R. N. Bergman, M. Uda, J. Tuomilehto, A. Cao, F. S. Collins, E. Lakatta, G. M. Lathrop, M. Boehnke, D. Schlessinger, K. L. Mohlke,

and G. R. Abecasis. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, 40(2):161–9, 2008.

CHAPTER

8

BMARS APPENDIX

8.0.1 Data Overview

Initial analysis of the dataset showed that there were a maximum of 126 (4.3%) genotypes missing over any one SNP, or a maximum of 116(9.4%) missing genotypes by individual. There were 368 SNPs (30%) with no missing genotypes and 1067(36.5%) individuals with no missing genotypes. The plots below show the distribution of missing genotypes.

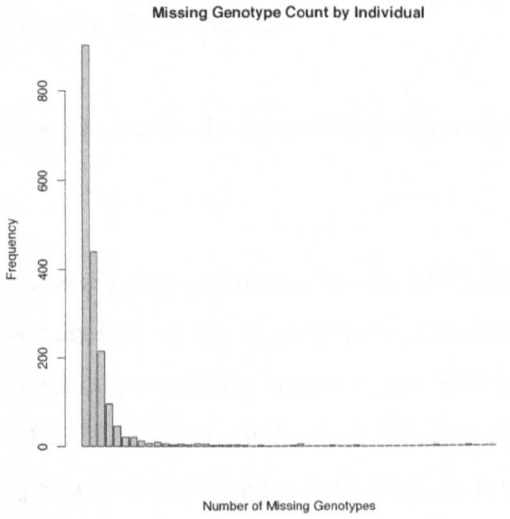


Figure 8.1: Missing genotype count by individual

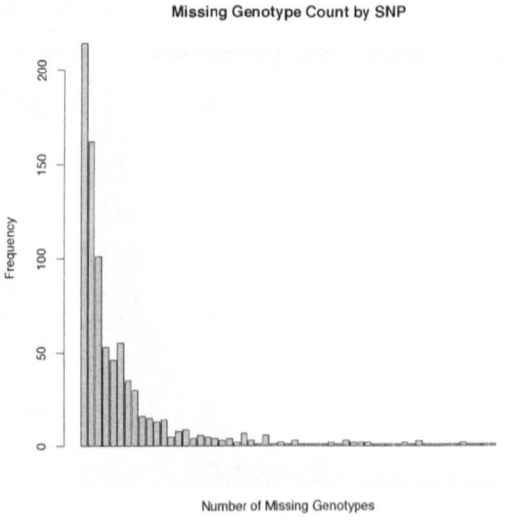


Figure 8.2: Missing genotype count by SNP

Figure8.1 shows that the majority of individuals who have any missing genotypes at all (approximately 900) are missing only 1 genotype out of 1230 SNPS in the dataset. As the number of genotypes missing increases, the frequency of individuals rapidly decreases. One individual has the maximum of 116 missing genotypes.

Figure 8.2 shows the number of individuals missing for each loci. There are 214 loci with 1 individual missing a genotype. Again, as the number of individuals missing genotypes increases for each loci, the frequency of occurrence decreases. Only one

loci has genotypes missing for 126 individuals.

Analysis of missing genotypes within the Spanish data showed similar levels of missing data.

The Spanish dataset contained genotype information for 813 individuals on 5375 SNPs on chromosome 6. Initial analysis of the dataset showed that there were a maximum of 42 out of 813 (5.2%) genotypes missing over any one SNP, or a maximum of 141 missing genotypes out of 5375 (2.6%) by individual. There were 3304 (61.5%) SNPs with no missing genotypes and 4 (0.5%)individuals with no missing genotypes. The plots below show the distribution of missing genotypes.

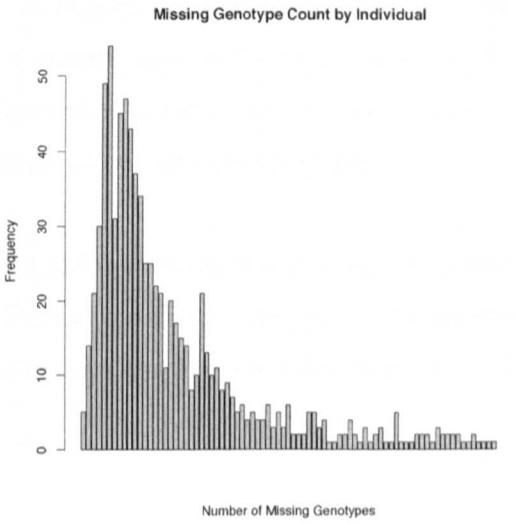


Figure 8.3: Missing genotype count by individual

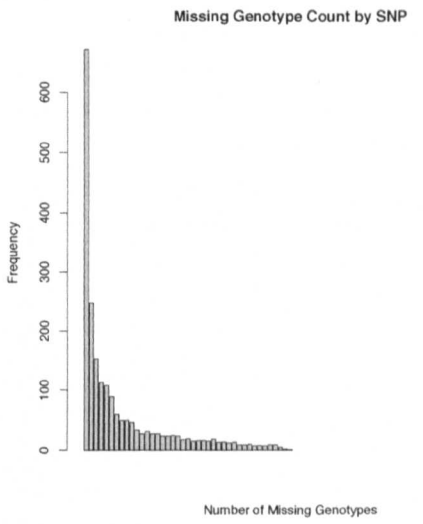


Figure 8.4: Missing genotype count by SNP

Figure 8.3 shows that the majority of individuals who have any missing genotypes at all (99.5% of them) are missing approximately 10 out of 5375 SNPs. As the number of genotypes missing increases, the frequency of individuals again decreases. One individual has the maximum of 141 missing genotypes.

Figure 8.4 shows the number of individuals with a missing genotype at each loci. There are 671 loci with 1 individual missing a genotype. As the number of individuals missing genotypes by loci increases, the frequency decreases. 1 loci has genotypes missing at 42 individuals.

8.1 LD Plots

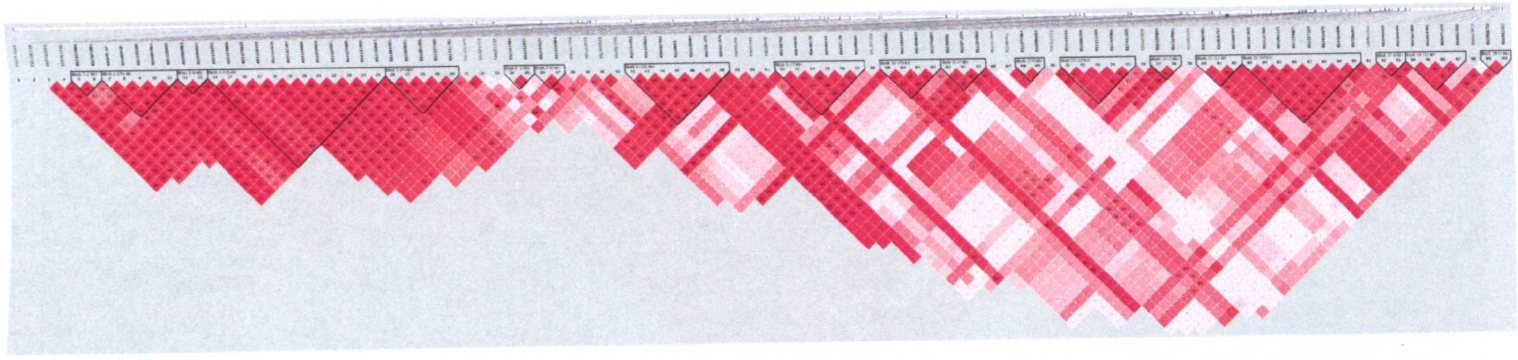


Figure 8.5: Linkage Disequilibrium of Top 100 UK/US SNPs

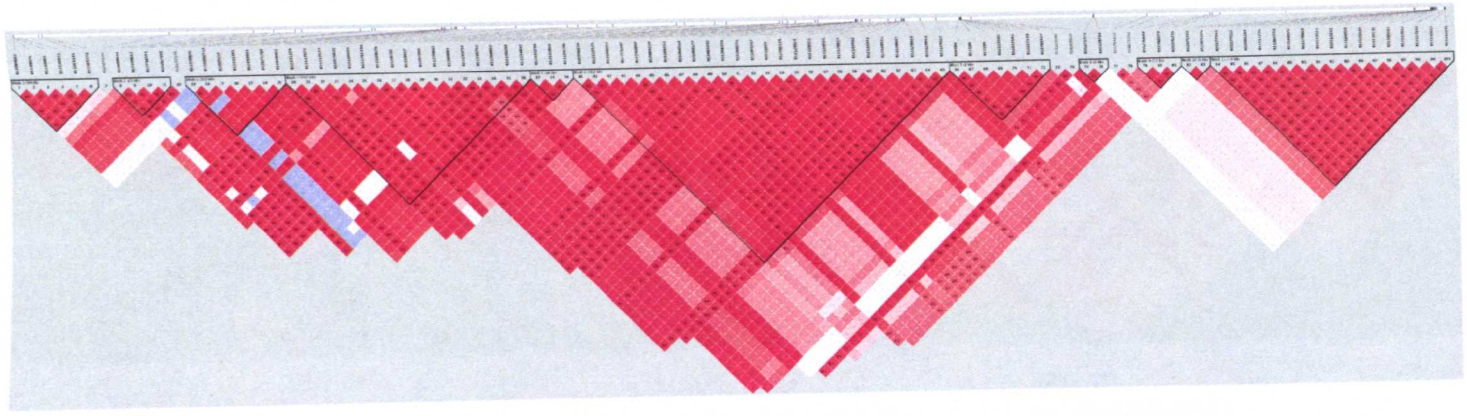


Figure 8.6: Linkage Disequilibrium of Top 100 Spanish SNPs

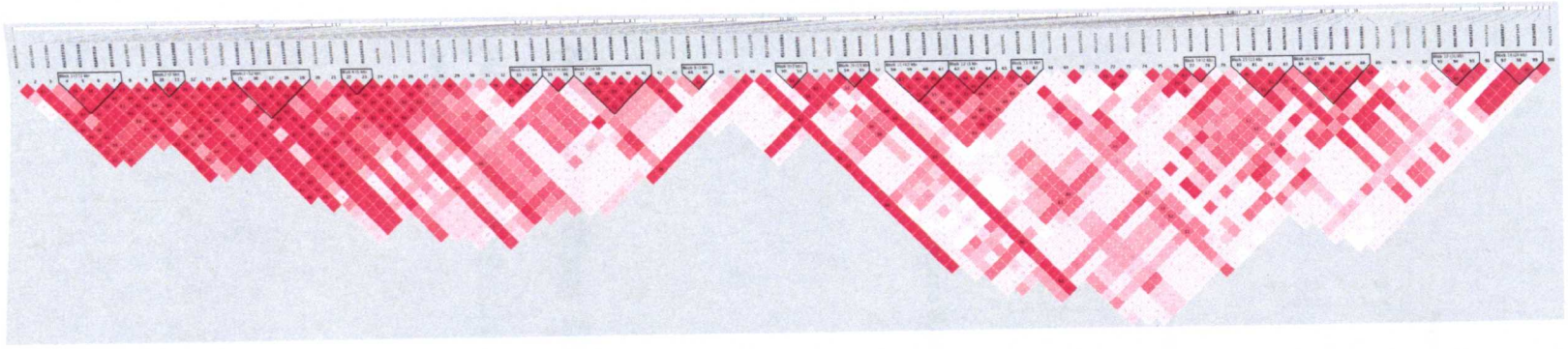


Figure 8.7: Linkage Disequilibrium of Top 100 Combined UK/US and Spanish SNPs

8.2 Posterior Probability Plots of UK vs US Data

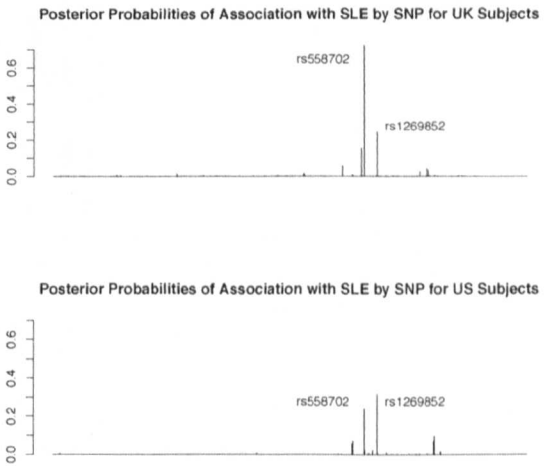
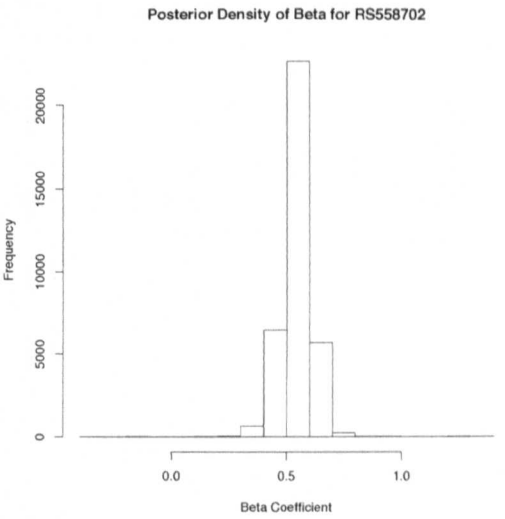
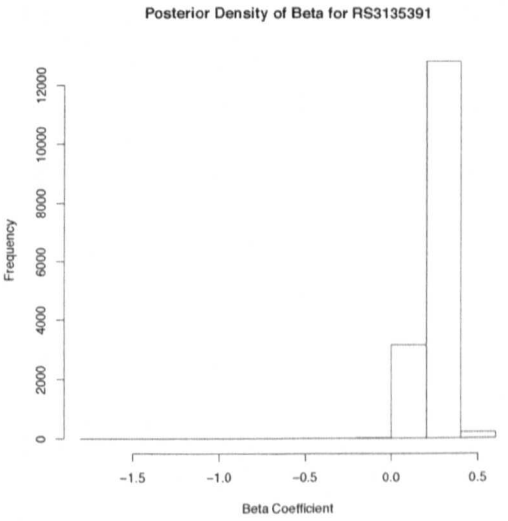


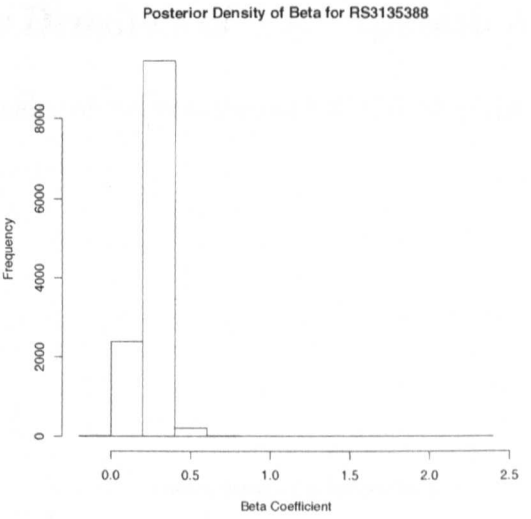
Figure 8.8: Posterior Probabilities by UK and US Datasets

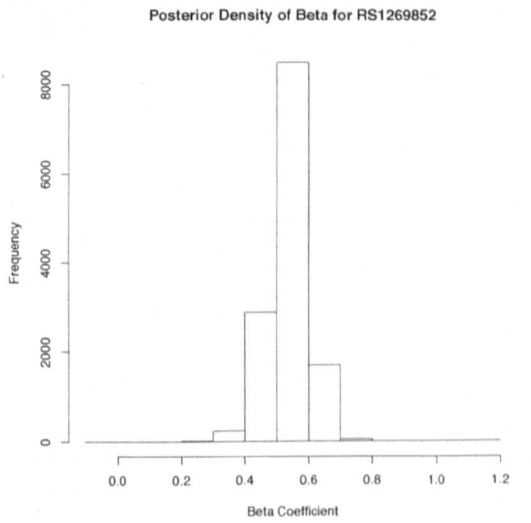
8.3 Posterior Densities of β for UK/US Analysis

Plots of posterior densities of β coefficients of top SNPs given they are in the UK/US model


Posterior Density of Beta for RS558702
Frequency
0
5000
10000
15000
20000
0.0
0.5
1.0
Beta Coefficient

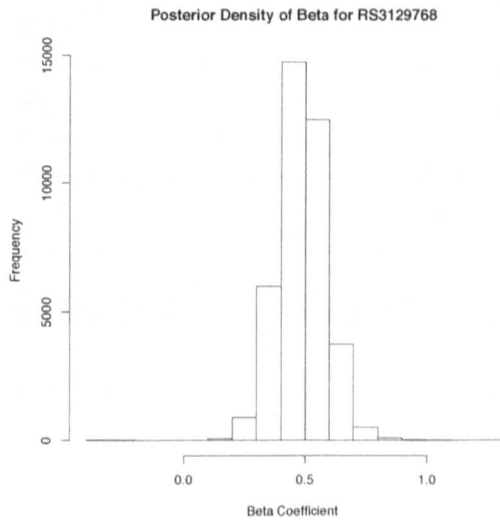

Posterior Density of Beta for RS3135391
Frequency
0
2000
4000
6000
8000
10000
12000
−1.5
−1.0
−0.5
0.0
0.5
Beta Coefficient


Posterior Density of Beta for RS3135388
Frequency
0
2000
4000
6000
8000
0.0
0.5
1.0
1.5
2.0
2.5
Beta Coefficient


Posterior Density of Beta for RS1269852
Frequency
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Beta Coefficient

8.4 Posterior Densities of β for Spanish Analysis

Plots of posterior densities of β coefficients of RS3129768 given it is in the model



8.5 Convergence Plots of Combined UK/US and Spanish Datasets Model

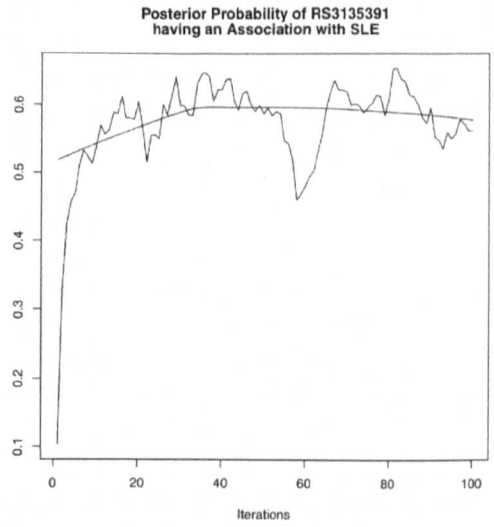


Figure 8.9: Convergence Plot of Posterior Probability of RS3135391 having an Association with SLE

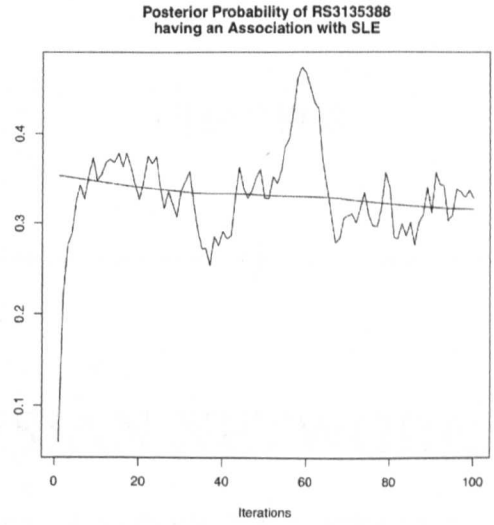


Figure 8.10: Convergence Plot of Posterior Probability of RS3135388 having an Association with SLE

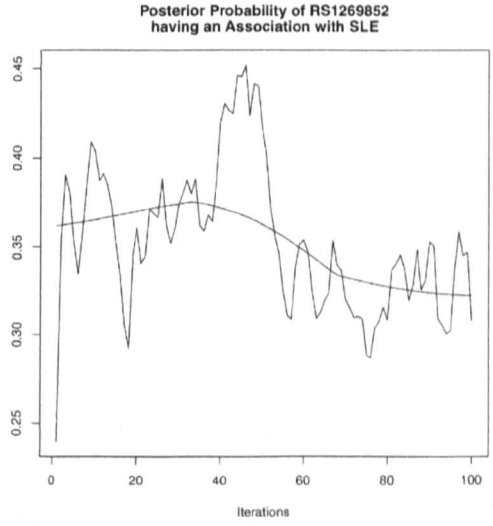


Figure 8.11: Convergence Plot of Posterior Probability of RS1269852 having an Association with SLE

CHAPTER

9

BAYESIAN NETWORKS FOR GENETIC ASSOCIATION STUDIES APPENDIX

9.1 Convergence Plots of Model 1

All plots are for first 100,000 iterations with a thin of 100.

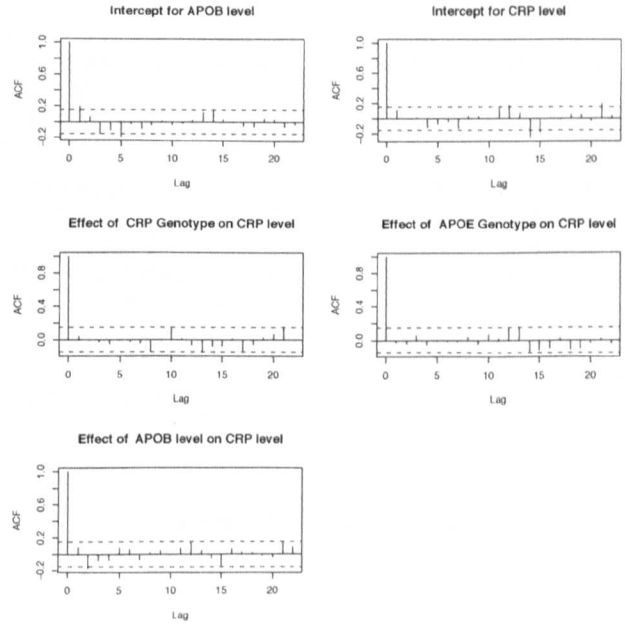


Figure 9.1: Autocorrelation function

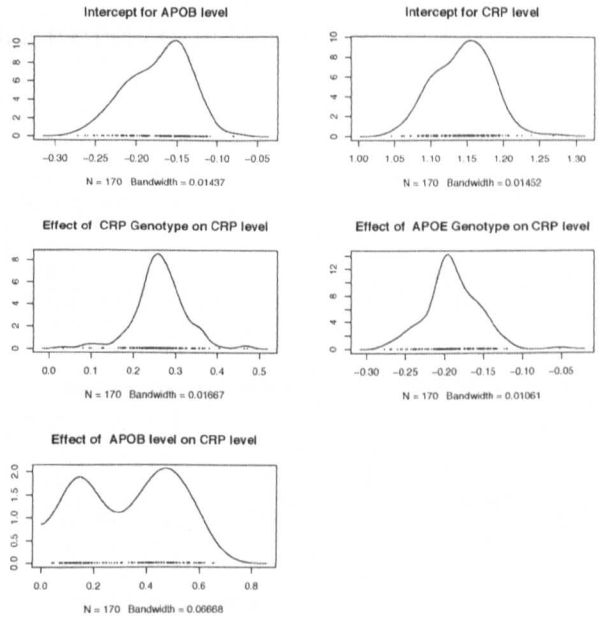


Figure 9.2: Density plots

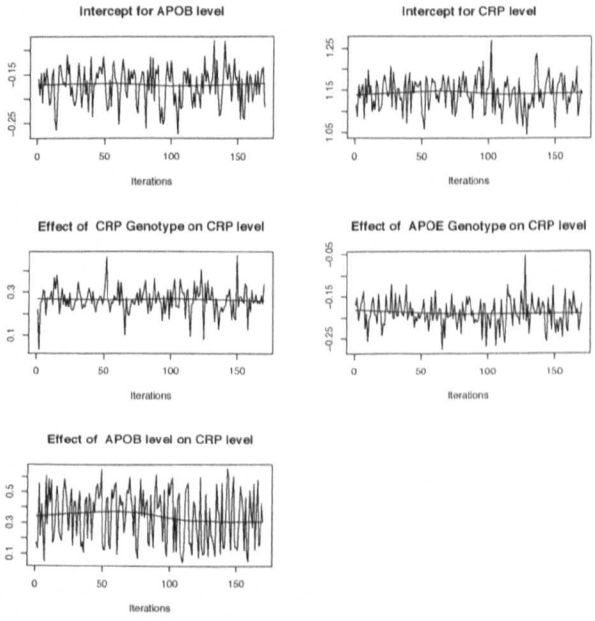


Figure 9.3: Trace plots

9.2 Convergence Plots of Model 2

All plots are for first 100,000 iterations with a thin of 100.

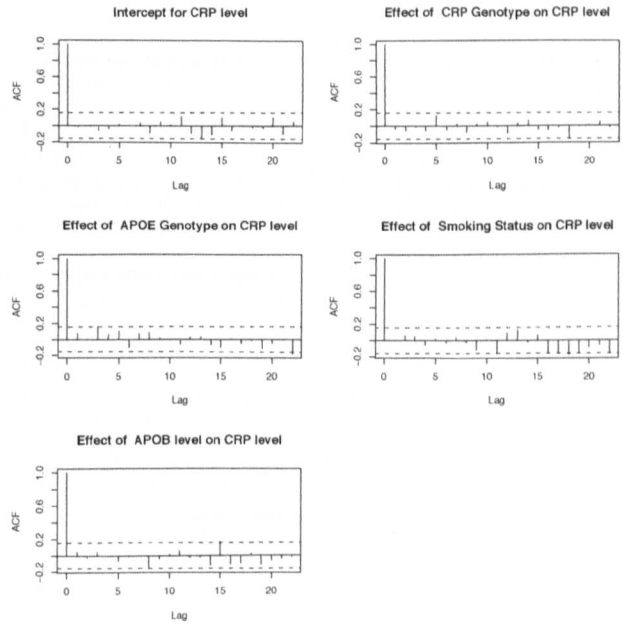


Figure 9.4: Autocorrelation function

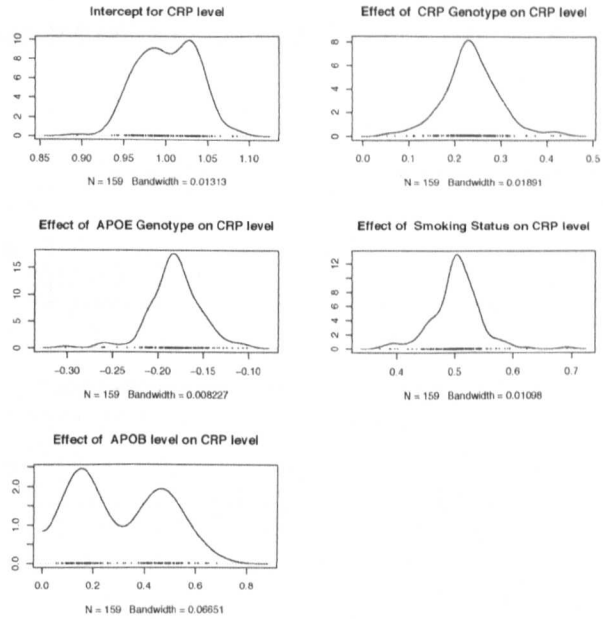


Figure 9.5: Density plots

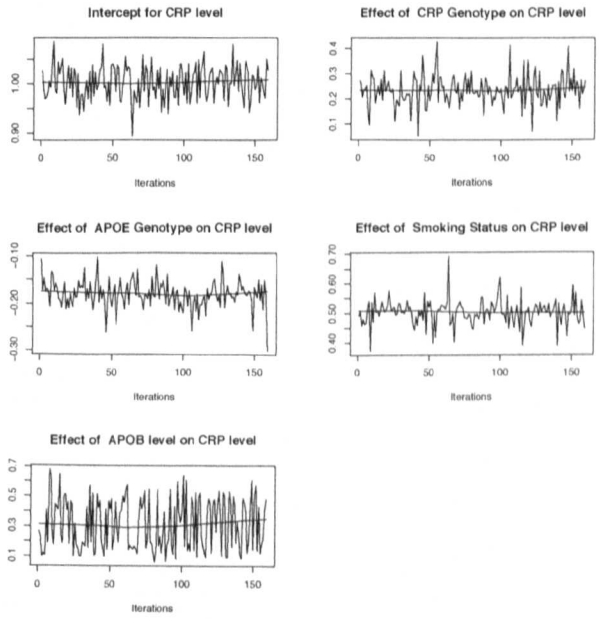


Figure 9.6: Trace plots

9.3 Convergence Plots of Model 3

All plots are for first 100,000 iterations with a thin of 100.

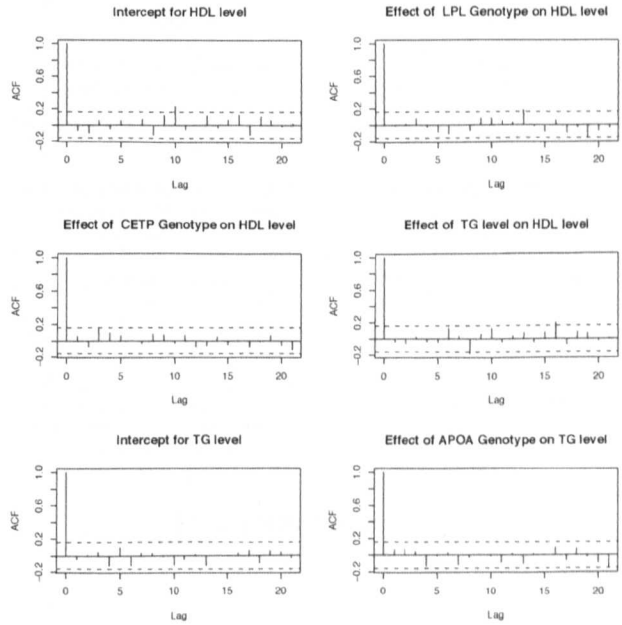


Figure 9.7: Autocorrelation function

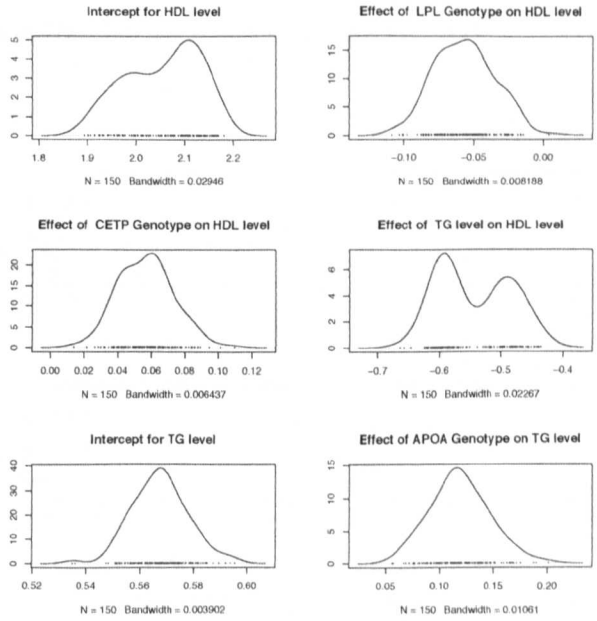


Figure 9.8: Density plots

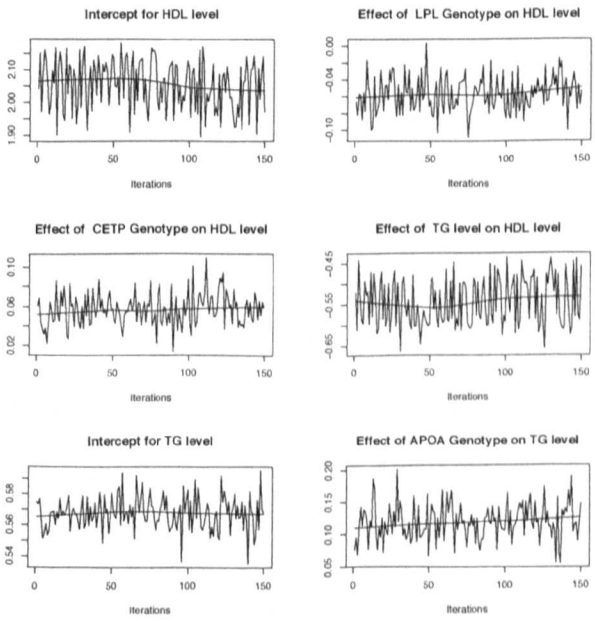


Figure 9.9: Trace plots

9.4 Convergence Plots of Model 4

All plots are for first 100,000 iterations with a thin of 100.

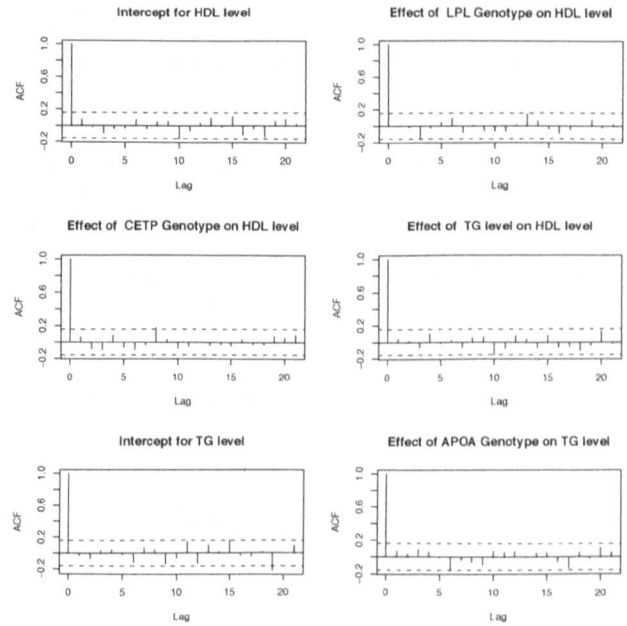


Figure 9.10: Autocorrelation function

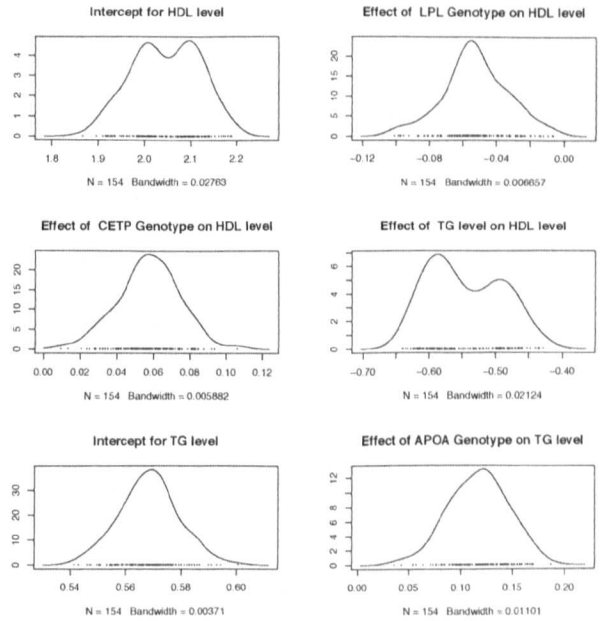


Figure 9.11: Density plots

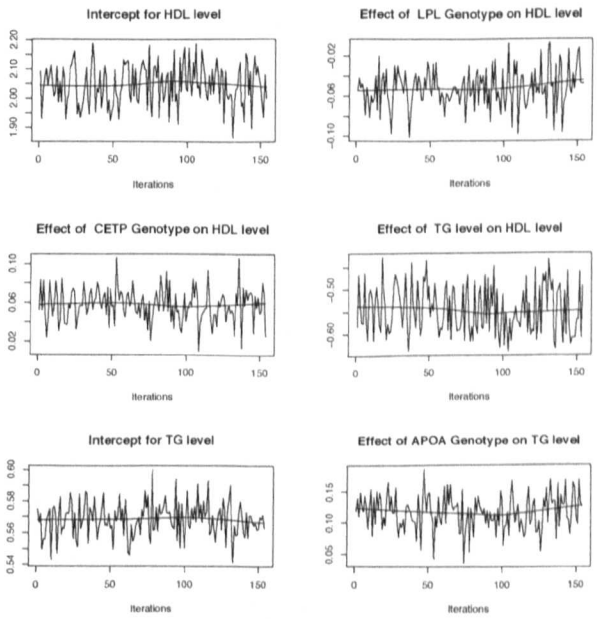


Figure 9.12: Trace plots